DOCUMENT RESUME

ED 328 584 TM 016 093

AUTHOR Shealy, Robin; Stout, William
TITLE An Item Response Theory Model for Test Bias.
INSTITUTION Illinois Univ., Urbana. Dept. of Statistics.
SPONS AGENCY Office of Naval Research, Arlington, VA. Cognitive and Neural Sciences Div.
PUB DATE 10 Jan 91
CONTRACT ONR-N00014-90-J-1940
NOTE 55p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Ability; Equations (Mathematics); Individual Differences; Item Bias; *Item Response Theory; *Mathematical Models; *Nonparametric Statistics; Scores; *Standardized Tests; *Test Bias

ABSTRACT

This paper presents a conceptualization of test bias for standardized ability tests which is based on multidimensional, non-parametric, item response theory. An explanation of how individually-biased items can combine through a test score to produce test bias is provided. It is contended that bias, although expressed at the item level, should be studied at the test level. The model postulates an intended-to-be-measured target ability and nuisance determinants whose differing ability distributions across examinee group cause bias. Multiple nuisance determinants can produce item bias cancellations, resulting in little or no test bias. Detection of test bias requires a valid subtest, whose items measure only target ability. A long-test viewpoint of bias is also developed. Four graphs and a 30-item list of references are included. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a REPORT SECURITY CLASSIFICATION: Unclassified	1b RESTRICTIVE MARKINGS:
2a SECURITY CLASSIFICATION AUTHORITY:	3 DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited.
2b DECLASSIFICATION/DOWNGRADING SCHEDULE:	
4 PERFORMING ORGANIZATION REPORT NUMBER(S): 1991-2	5 MONITORING ORGANIZATION REPORT NUMBER(S):
6a NAME OF PERFORMING ORGANIZATION: University of Illinois Department of Statistics; 6b OFFICE SYMBOL (If applicable):	7a NAME OF MONITORING ORGANIZATION: Cognitive Science Program Office of Naval Research (Code 1142 CS)
6c ADDRESS (City, State, and ZIP Code): 101 Illini Hall 725 S. Wright Street Champaign, IL 61820	7b ADDRESS (City, State, and ZIP Code): 800 N. Quincy Street Arlington, VA 22217-5000
8a NAME OF FUNDING/SPONSORING ORGANIZATION: ; 8b OFFICE SYMBOL (If applicable):	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: N00014-90-J-1940
8c ADDRESS (City, State, and ZIP Code): 101 Illini Hall 725 S. Wright Street Champaign, IL 61820	10 SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO: 61153N; PROJECT NO: RR04204; TASK NO: RR04204-01; WORK UNIT ACCESSION NO: 4421-548

11 TITLE (Include Security Classification): An Item Response Theory Model for Test Bias

12 PERSONAL AUTHOR(S): Robin Shealy and William Stout

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
technical	FROM 1987 TO 1990	January 10, 1991	49

16 SUPPLEMENTARY NOTATION: To appear in Differential Item Functioning, Theory and Practice, L. Erlbaum, 1992

17 COSATI CODES			18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	See reverse

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

See reverse

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	
22a NAME OF RESPONSIBLE INDIVIDUAL: Dr. Charles E. Davis	22b TELEPHONE (Include Area Code): (703) 696-4046; 22c OFFICE SYMBOL: ONR-1142CS

DD Form 1473, JUN 86 — Previous editions are obsolete — SECURITY CLASSIFICATION OF THIS PAGE

S/N 0102-LF-014-6603

Abstract

A multidimensional non-parametric IRT model of *test bias* is presented, providing an explanation of how individually-biased items can combine through a test score to produce test bias. The claim is thus that bias, though expressed at the item level, should be studied at the test level. The model postulates an intended-to-be-measured *target ability* and nuisance determinants whose differing ability distributions across examinee group cause bias. Multiple nuisance determinants can produce *item bias cancellation*, resulting in little or no test bias. Detection of test bias requires a *valid subtest*, whose items measure only target ability. A long-test viewpoint of bias is also developed.

Keywords: latent trait theory, item response theory, item bias, test bias, DIF, long-test theory, essential unidimensionality, item bias cancellation, target ability, nuisance determinants, valid subtest.

1 Introduction

The purpose of this paper is to present an Item Response Theory (IRT) based conceptualization of test bias for standardized ability tests. By "test bias" we mean a formalization of the intuitive idea that a test is less valid for one group of examinees than for another group in its attempt to assess examinee differences in a prescribed latent trait, such as mathematics ability. It will be seen that test bias is the result of individually-biased items acting in concert through a test scoring method, such as number correct, to produce a biased test. In a subsequent paper of ours, this new conceptualization of test bias is used to undergird a new statistical test for psychological test bias (Shealy and Stout, 1990). Also, a large-scale simulation study (Shealy, 1989) has been conducted of the performance properties of this statistical procedure, in particular as compared with the Holland and Thayer (1988) modification of the Mantel-Haenzel test.

We mention three distinct features of the conceptualization of bias presented herein. First, it provides a mechanism for explaining how several individually-biased items can combine through a test score to exhibit a coherent and major biasing influence at the test level. In particular, this can be true even if each individual item displays only a minor amount of item bias. For example, "word problems" on a "mathematics test" that are too dependent on sophisticated written English comprehension could *combine* to produce pervasive test bias against English-as-a-second-language examinees. A second feature, possible because of our multidimensional modeling approach, is that the underlying mechanism that produces bias is addressed. This mechanism lies in the distinction made between the ability the test is intended to measure, called the *target ability*, and other abilities influencing test performance that the test does not intend to measure, called *nuisance determinants*. Test bias will be seen to occur because of the presence of nuisance determinants possessed in differing amounts by different examinee groups. Through the presence of these nuisance determinants, bias then is expressed in one or more items. A third feature, also possible because of our multidimensional modeling approach, is that a careful distinction is made between genuine test bias and non-bias differences in examinee group performance that are caused by examinee group differences in target ability distributions. It is important that the latter not be mistakenly labeled as test bias.

The novelty of our approach to bias lies not so much with its recognition of the role of nuisance determinants in the expression of test bias, but rather in the explicit multidimensional IRT modeling

of bias, which in turn promises a clear and thorough understanding of bias.

2 An Informal Description of Test Bias

We begin with an informal definition of test bias.

Definition 2.1. *Test bias occurs if the test under consideration is measuring a quantity in addition to the one the test was designed to measure, a quantity that both groups do not possess equally.* □

It is important to note that this notion of test bias grows out of the traditional non-IRT notion of test bias based on differential predictive validity. Papers by Stanley and Porter (1967), Temp (1971), and in particular Cleary (1968), exemplify this classical predictive view of bias. These studies used standardized tests to predict *performance* on a *particular task*; if the predictive link from test to task was different for the two studied groups, then test bias was suspected. Cleary (1968), in a seminal paper on test bias, studied bias in the prediction of college success of black and white students in integrated colleges, using SAT verbal and math scores. Her intent was to determine if the expected first year GPA (grade point average) for Whites was different from that for Blacks, after the two groups had been *matched* on SAT score; hence, the linear regression of first year GPA on SAT verbal (or math) score was separately fit for both groups and compared. If the expected criterion (GPA) for those examinees attaining a particular test score (e.g., SAT combined score) were different across group, the test score was considered a biased predictor of performance and test bias was deemed to be present. The purpose of the Cleary study was *predictive*; the regressions of *criterion* on *test score* therein were compared across group to see if the test score equitably predicts the performance measured by the criterion.

Our focus shifts hereafter to regressing test score on criterion. The purpose of the reversed regression is to *corroborate* that the prediction of a criterion by a test is equitable across group, thereby exposing the conceptual underpinning for IRT modeling of test bias – in particular our modeling of test bias. The regressions of *test score* on *criterion* are compared to answer the following question: are the average test scores for both groups the same after the groups have been matched on criterion performance?

This shift to a corroborative point of view brings us again to the informal definition above. The difference across group in the regressions of test score on criterion (other than that caused by statistical error) is due to an undesirable causative factor *other* than the criterion; that is, at

least some of the test questions must be measuring something in addition to what the criterion measures. Furthermore, this difference is due only to the undesirable factor, because the criterion has been held equal in the two groups.

If in addition to the reversal of the regression variables just described, the criterion is now chosen as internal to the test instead of external to it, the concept of the internal assessment of bias results. This internal criteria becomes the "yardstick" by which the test is judged biased or not; it is *a portion of the test itself.* The implicit assumption is that the "yardstick" portion of the test consists of items known to measure only what they are supposed to be measure.

An example adapted from Shepard (1982) clearly illustrates this internally-assessed bias: a verbal analogies test is used to compare reasoning abilities of German and Italian immigrants to the United States, the two populations matched on English fluency. However, 20% of the test items are based on words with Latin origins, whereas the remainder have linguistic structure equally familiar to both groups. Here, the items with Latin origin words are possibly biased. A reasonable internal criterion with which to assess this bias would be a score based on the responses to the linguistically neutral items; for, it is assumed that these items are validly measuring what the test is intended to measure.

The internal assessment viewpoint of test bias can be clarified by noting two distinctions between it and the classical test theory based differential regression conceptualization of Cleary and others:

(A) The "yardstick" (criterion), which was a measurement of task performance (e.g., 1st year GPA in the Cleary study), is now a score internal to the test (e.g., score on the linguistically neutral items in the Shepard example). This internal criterion is most often an aggregate measure of a portion of the test item responses (typically number right).

(B) The differential regression approach used regressions of the external *criterion* on *test score* in a *predictive* context. In internally-assessed bias studies, the responses of one or more items suspected of bias are regressed on the internal criterion as a *corroborative* statistical test that these "suspect" items are measuring the same thing that the internal criterion is measuring.

This brings us to an essential question: what is the internal criterion measuring? It is measuring a theoretically postulated psychometric construct that is intended to be generalizable to a variety of possible future tasks; i.e., a latent ability of an IRT model. Thus IRT modelling of

bias becomes appropriate. An example will illustrate: the SAT math test is designed to measure a construct, "mathematical ability", which is intended to predict an examinee's future success in a set of quantitatively-oriented college courses that require a component of such ability. Rather than assessing the SAT test against the corresponding set of criterion measures of performance in these courses, we wish to assess the test against the construct itself; to do so we turn to the test itself to verify that the proper measurement of mathematical ability is being done. The *internal criterion* measures this ability construct, and internal test bias is defined with respect to this construct.

The generalizability of performance measurements on a variety of tasks to a single construct, as described above, provides one motivation to shift to internally assessed bias studies. An additional motivation is the practice in recent years of creating *item pools*, large numbers of items that are to be used in forming multiple versions of a standardized test (see, for example Hambleton and Swaminathan, 1985, Ch. 12). A newly constructed set of items intended for inclusion in the item pool can be tested for bias, relative to the ability construct that the pool is supposedly measuring, by employing internal bias detection techniques.

Internally assessed bias studies with a variety of test populations have been done: Cotter and Berk (1981) attempted to detect bias in the WISC-R test with white and minority children. Dorans and Kulick (1983), in a series of studies done at Educational Testing Service, study the possible effect of differential mastery of written English between native born Americans and English-as-a-second-language Oriental students on scores of selected items on a mathematics "word problem" test.

Item bias studies such as the ones above usually focus on single item at a time; if several items in these studies are simultaneously found to be biased, it is a result of statistical bias procedures conducted for each item separately, which raises delicate questions about simultaneous statistical inference. Moreover, in a modeling sense, no causative reasons for the observed simultaneous bias are explored by item bias studies. This paper studies a form of *test bias* relative to an internal criterion; this kind of test bias considers the set of test items acting as a unit (via a common causal mechanism) and combining through a test scoring method. The precise formulation of test bias and a contrast of it to item bias is presented in Section 4.

We now consider the question of test bias relative to an internal criterion more carefully. Consider a situation where a single verbal analogy item is embedded in two different tests, tests $\mathcal{M}$ and

$\mathcal{V}$ say. Test $\mathcal{V}$ is composed of verbal analogy items, as intended, and Test $\mathcal{M}$ consists of mathematics calculation items, as intended, except for the single embedded item. Assume that each item in Test $\mathcal{V}$ does not contain any culture-dependent material that may favor one group. The embedded verbal item is *not* biased in Test $\mathcal{V}$, but the potential for bias of this item is large in Test $\mathcal{M}$, because the item measures something other than the intended-to-be-measured mathematical calculation skill. This illustrates a key component of test bias, aptly stated by Mellenbergh (1983, p. 294): "An item can be biased in one set of items, whereas it is unbiased in another set." Shepard (1982) also points out this relativity feature: "... if a test of spatial reasoning inadvertently included several vocabulary items, they would be biased indicators of the [ability being measured]" and "... in a test composed equally of two types of items reflecting... two different [ability] constructs, it will be a dead heat to decide statistically which set defines the test [ability] and which set becomes a biased measure of it."

Implicit in the above discussion is the assumption that a portion of the test defines the internal criterion by which the remainder is measured for the presence of bias. A collection of items defining the internal criterion will be called a *valid subtest*. An informal definition of a valid subtest can now be given: A subtest is valid with respect to a specified "target" ability if the subtest score is judged to be *measuring only the intended target test ability*, i.e., it stands as a "proxy" of the ability one intends to measure. More precisely, if all of the items of the subtest measure *only* the intended ability then the subtest is said to be valid.

There is a point about this definition that needs mentioning. Primarily, the existence and identification of a valid subtest is an empirical decision based on expert opinion or data at least in part external to the data set in question. Subtest validity *cannot* be established based on the test data set alone nor can it be theoretically deduced. The "burden of proof" is an empirical one and lies with the test constructor. If all the items of a test depend on a second determinant (for example, if the responses to all items depend on familiarity with standardized tests) then a valid subtest will not exist. Note that this is true even if the two groups are not differentially penalized by this dependence of test items on familiarity with standardized tests. Thus, the actual presence of test bias is logically independent of the existence of a valid subtest to be used for the assessment of test bias.

In our framework, it must be assumed that there is a valid subtest if we are to internally detect

test bias; otherwise, it is intrinsically nondetectable internally. The responses to the valid subtest are used to tackle the central problem in the identification of test bias: the need to distinguish between group differences attributable to the ability construct intended to be measured and that due to unwanted ability determinants. Because the valid subtest is assumed to measure only the desired ability, then no measures external to the test are required to assess that ability, although to improve accuracy it may be beneficial to also use external data, especially if the valid subtest is short or if the assumption of its validity is at all suspect. Matching examinees using a valid subtest score controls for group differences in the intended-to-be-measured ability and isolates differences due to the unwanted determinants. A more rigorous formulation of "valid subtest" is set out in Section 4.

In these discussions of test bias relative to an internal criterion, multidimensionality has implicitly been invoked; it is impossible to discuss test bias without invoking it. The informal definition of test bias stated above employs multidimensionality: there is mention of the quantity the "test was designed to measure" and one "in addition to" this quantity. Lord (1980, p. 220) recognized this in his discussion of item bias: "if many of the items [in a test] are found to be seriously biased, it appears that the items are not strictly unidimensional".

Bias in one or more items has typically been attributed to special knowledge, unintended to be measured, that is more accessible to one of the test-taking groups. Ironson, Homan, Willis and Signer (1984) performed a bias study that involved planting within a mathematics test mathematics word problems that required an extremely high reading level to solve them. They state their conclusion that "... bias is sometimes thought of as a kind of multidimensionality involving measurement of a primary dimension and a second confounding dimension". Our viewpoint here is that bias is *always* the result of multidimensionality.

The "primary dimension" is referred to in this paper as *target ability*, because this is the ability the test intends to measure. The "confounding dimension" is referred to as a *nuisance determinant*. In the Shepard verbal analogies example above, the target ability is reasoning ability, which 80% of the items solely measure, while the nuisance determinant is *familiarity with Latin linguistic roots*.

The full formulation of test bias is set out in Section 4–it involves certain subtleties not discussed here. The group differences in ability level of a latent nuisance determinant provide a common causative mechanism for bias in any collection of items on a test contaminated with such

a determinant. This is the single most important conceptual difference between the test bias model developed in this paper and previous item bias work: the existence of a postulated common *latent* cause for the *manifestation* of bias across a group of test items.

3 The IRT Model for Test Responses

Herein we present the nonparametric multidimensional IRT model underlying our modeling of test bias. Consider a group of $\mathcal{G}$ of examinees; the sample of examinees to take a test is considered to be drawn at random from this population. A test is simply a collection of items; a *test response* of length N is the corresponding set of responses, for a randomly-chosen examinee from $\mathcal{G}$, denoted by

$$\underline{U} = (U_1, \ldots, U_N) \tag{3-1}$$

where the U_i are random variables taking on

$$U_i = \begin{cases} 0 & \text{if response to item } i \text{ is incorrect;} \\ 1 & \text{if response is correct.} \end{cases}$$

The IRT model is composed of two components that generate $\underline{U}$: (1) a d-dimensional examinee ability parameter and (2) a set of item responses functions (IRFs), one for each item, which determine the probability of correct response for the items. The IRT model is usually conceived as a *unidimensional* ($d = 1$) model; here, a multidimensional ($d > 1$) model will be presumed.

Let us now further set notation. The ability vector is

$$\underline{\theta} = (\theta_1, \ldots, \theta_d) \tag{3-2}$$

for an arbitrary examinee from $\mathcal{G}$. A distribution of $\underline{\theta}$ over $\mathcal{G}$ is induced by choosing examinees at random from $\mathcal{G}$; the multivariate random variable is designated

$$\underline{\Theta} = (\Theta_1, \ldots, \Theta_d) \tag{3-3}$$

Examinee independence is assumed; i.e., J examinees from $\mathcal{G}$ have ability parameters

$$\{\underline{\Theta}(j) : j = 1, \ldots, J\}$$

independent and identically distributed (iid) in j. Item i's IRF, which is interpreted as the probability that an examinee with ability $\underline{\theta}$ will answer item i correctly, is denoted:

$$P_i(\underline{\theta}) = P[U_i = 1 | \underline{\Theta} = \underline{\theta}] \equiv P[U_i = 1 | \underline{\theta}].$$

Our interpretation of $P_i(\underline{\theta})$ is the sampling one: among all examinees having ability $\underline{\theta}$, the expected proportion of them getting item i correct is $P_i(\underline{\theta})$.

The basic philosophy of the IRT model is that a *latent* distribution of abilities in a Group $\mathcal{G}$ drives the *manifest* distribution of item responses. The fundamental identity relating the responses $\underline{U}$ to the examinee group ability variable $\underline{\Theta}$ is

$$\begin{aligned} P[\underline{U} = \underline{u}] &= \int_{\underline{\theta}} P[\underline{U} = \underline{u} | \underline{\Theta} = \underline{\theta}] dF(\underline{\theta}), \\ &\text{for all } \underline{u} = (u_1, \ldots, u_N), \qquad (\text{ each } u_i = 0 \text{ or } 1), \end{aligned} \tag{3-4}$$

where $F(\cdot)$ is the cumulative distribution function (cdf) of $\underline{\Theta}$. There are two fundamental assumptions on the conditional test response probability $P[\underline{U} = \underline{u} \mid \underline{\theta}] \equiv P[\underline{U} = \underline{u} \mid \underline{\Theta} = \underline{\theta}]$ usually assumed in IRT modeling. To introduce these, recall two standard definitions about ordering in d-dimensional Euclidean space: (i) Let $\underline{z}$ and $\underline{z}'$ be vectors. Then $\underline{z} < \underline{z}'$ if $z_i \leq z_i'$ for $i = 1, \ldots d$ and for at least one i, $z_i < z_i'$. (ii) Let $\underline{z}$ and $\underline{z}'$ be vectors. The real valued function $f(\underline{z})$ is *strictly monotone* if for any $\underline{z} < \underline{z}'$, $f(\underline{z}) < f(\underline{z}')$.

The fundamental IRT assumptions are:

Assumption 3.1. *Local independence in $\underline{\theta}$: for every $\underline{\theta}$,*

$$P[\underline{U} = \underline{u} | \underline{\theta}] = \prod_{i=1}^{N} P[U_i = u_i \mid \underline{\theta}] \quad \text{for all } u_i = 0 \text{ or } 1; i = 1, \ldots, N. \tag{3-5}$$

Assumption 3.2. *Strict monotonicity of IRFs: The item IRFs $\{P_i(\underline{\theta}) : i = 1, \ldots, N\}$ are strictly monotone in $\underline{\theta}$. That is, for any i, $P_i(\underline{\theta}') > P_i(\underline{\theta})$ if $\underline{\theta}' > \underline{\theta}$ in the sense of (i) above.*

It is convenient to combine (3-4) and Assumption 3.1 in the following manner:

$$P[\underline{U} = \underline{u}] = \int_{\underline{\theta}} \prod_{i=1}^{N} P_i(\underline{\theta})^{u_i} (1 - P_i(\underline{\theta}))^{1-u_i} dF(\underline{\theta}) \tag{3-6}$$

for all $\underline{u}$.

The notion of the dimensionality d of $\underline{U}$ can be mathematically formalized but for the purposes of this paper it is unnecessary to do so.

Definition 3.1. Let $\underline{U}$ be a test response as in (3-1). An *IRT representation of $\underline{U}$* is the structure

$$\{d, \underline{\Theta}, F(\underline{\theta}), \{P_i(\underline{\theta}) : \quad i = 1, \ldots, N\}\} \tag{3-7}$$

where (3-4), Assumption 3.1, and Assumption 3.2 hold. □

In this paper we often want to consider a test item's operating characteristic with respect to a specific single component of $\underline{\theta}$ (say θ_1). This is accomplished by "marginalizing out" the remaining components in the $\underline{\theta}$-vector from the item's IRF, resulting in the *marginal item response function* (marginal IRF). Conceptually, this IRF is a unidimensional reduction of the original one and can be considered as a unidimensional IRF for the unidimensional ability θ_1. The following definition is due to Stout (1989).

Definition 3.2. *Let $P(\underline{\theta})$ be an IRF. The marginal IRF $T(\theta_1)$ of $P(\underline{\theta})$ with respect to Θ_1 is defined by*

$$T(\theta_1) = E[P(\underline{\Theta})|\Theta_1 = \theta_1]. \qquad \square$$

The marginal IRF is essential in the discussion of modeling test bias in Section 4, where a single component θ_1 of $\underline{\theta}$ designated as the *target ability* will be considered. Because target ability is the ability the test designer desires to measure using the items, the marginal IRF with respect to this ability is a useful concept.

In order for $T(\theta_1)$ to be an IRF it must be strictly monotone; this does not follow for the marginal IRFs of a test from the assumptions of our IRT representation (3-7). However, very mild regularity conditions suffice to produce strict monotonicity, as has been shown by Stout (1989). To this end, we need the concept of stochastic ordering.

Definition 3.3. *Let $\underline{Z}$ be a random vector with distribution indexed by a parameter γ. $\underline{Z}$ is strictly stochastically increasing in γ if for every $\underline{z}$ in the range of $\underline{Z}$*

$$P[\underline{Z} > \underline{z}; \gamma] < P[\underline{Z} > \underline{z}; \gamma'] \text{ if } \gamma < \gamma'.$$

Strict monotonicity of the marginal IRF with respect to θ_1 follows under the reasonable assumption of stochastic order in Θ_1:

Theorem 3.1. *(See Stout, 1989). If $\underline{\Theta}|\Theta_1 = \theta_1$ is strictly stochastically increasing in θ_1 in the sense of Definition 3.3 and the IRF $P(\underline{\theta})$ is strictly monotone in $(\theta_2, \ldots, \theta_d)$ then the marginal IRF of $P(\underline{\theta})$ with respect to θ_1 is strictly monotonic.*

Remark. Note in Theorem 3.1 that $P(\underline{\theta})$ is not assumed to be strictly monotone in θ_1, the first component of $\underline{\theta} \equiv (\theta_1, \theta_2, \ldots, \theta_d)$.

A note on IRT model assumptions should be emphasized here. IRT models are commonly parameterized; that is, the IRFs and ability distribution are members of parametric families. Typical assumptions are that $\underline{\Theta}$ is unidimensional with a standard normal distribution and that a two or three parameter normal ogive model or a one, two, or three parameter logistic model is assumed for the IRFs. In this paper, we assume only that the IRFs $\{P_i(\underline{\theta})\}$ are continuous, with $\underline{\theta}$ usually multidimensional.

4 Test Bias in the IRT Model

In this section our multidimensional IRT based notion of *test bias* using the IRT model of Section 3 is developed. Section 4.1 provides a brief presentation on IRT *item bias* as currently usually defined in the psychometric literature. Section 4.2 sets up the multidimensional IRT framework for test bias modeling; *target ability* and *nuisance determinants* are defined. Section 4.3 develops test bias in terms of its components: *potential* for bias, *expressed* bias, and the combining of expressed item biases through a test scoring method. Section 4.4 considers item bias cancellation when the nuisance determinants are multidimensional. Finally, Section 4.5 formally considers the notion of a valid subtest.

4.1 Existing IRT Item Bias Definition

In this section the concept of IRT-modeled item bias (in some contexts called DIF, for differential item functioning) currently in widespread use is presented as a backdrop for the development of multiple-item test bias, which is treated in Sections 4.2 and 4.3. An item is biased, according to Hambleton and Swaminathan, (1985, p. 285) if its (necessarily unidimensional) item response functions across groups are not identical. A formal definition is given below.

Definition 4.1. *Item bias. Let two groups of examinees be indexed by $g = 1, 2$. For each g, denote*

$$\underline{U}_g = (U_{1g}, \ldots, U_{Ng}) \tag{4-1}$$

to be the test response from an N-item test for a randomly chosen examinee from Group g. Assume that a unidimensional IRT model fits each group, with IRT representation for $\{\underline{U}_g; g = 1, 2\}$ (recall Definition 3.1):

$$\{d = 1, \Theta_g, F_g(\theta), \{P_j(\theta): \quad j = 1, \ldots, i-1, i+1, \ldots, N; P_{ig}(\theta)\}, g = 1, 2\}, \tag{4-2}$$

where $F_g(\theta)$ denotes the cdf of Θ_g. (Note, as the subscript notation indicates, that all items except the ith item have group invariant IRFs while item i has an IRF that possibly differs for the two groups.)

(i) Item bias occurs in item i at θ if the group specific probabilities of correct response at θ are different; i.e., the group IRFs are different at θ:

$$P_{i1}(\theta) \equiv P[U_{i1} = 1 \mid \Theta_1 = \theta] \neq P[U_{i2} = 1 | \Theta_2 = \theta] \equiv P_{i2}(\theta).$$

(ii) Item bias occurs in item i if there exists some value θ for which item bias occurs at θ. □

It is important to observe that the "bias" of item i is defined relative to the other $N - 1$ items, which are assumed invariant and hence "unbiased" with respect to the two groups.

Item bias models have traditionally been parametric. Wright, Mead and Draba (1976) and Holland and Thayer (1988) consider a biased item generated by Rasch IRFs with the IRF difficulties (b's) different for the 2 groups. The more general 2PL and 3PL models, with different discriminations (a's) and guessing parameters (c's) across group, have been studied by Hulin, Drasgow and Komocar (1982), Linn, Levine, Hastings, and Wardrop (1981), and Thissen, Steinberg, and Wainer (1988), among many others.

Item bias addresses differential performance across group for a single item at a time. If several items display bias relative to the remaining assumed group invariant items according to Definition 4.1– modified to allow several IRFs to possibly differ across group–there are no components in Definition 4.1 that provide the facility to explain simultaneous item biasing due to a single underlying reason. This provides the motivation for an IRT framework that explains such pervasiveness of item bias.

4.2 The IRT Framework for Multidimensional Test Bias

In our treatment, test bias is modeled using the nonparametric multidimensional IRT framework described in Section 3. The multidimensionality of the underlying latent abilities for the two groups provides the environment from which bias expresses itself in one or more items. A crucial component in this test bias model is the modeling of a *pervasive* nuisance determinant, which contaminates a significant portion of the test items. This modeling viewpoint is an attempt to retain the view that

bias originates at the test question level yet to allow for the possibility of bias expressed through a test score as in the classical differential regression approach discussed above in Section 2.

The setup of the multidimensional IRT model for a test administration to two groups is as follows. The IRT representation (3-7) is assumed to hold for the combined two-group population of examinees. This representation induces a separate IRT representation of the form of (3-7) for each of the two groups:

$$\{d, \underline{\Theta}_g, F_g(\underline{\theta}), \{P_{ig}(\underline{\theta}): \quad i = 1, \ldots, N\}\}, g = 1, 2, \tag{4-3}$$

where $\underline{\Theta}_g$ here denotes $\underline{\Theta}$ restricted to Group g, $F_g(\underline{\theta})$ denotes the cdf of $\underline{\Theta}_g$, and $P_{ig}(\underline{\theta})$ denotes the ith IRF for a randomly selected examinee from the subpopulation of Group g examinees of ability $\underline{\theta}$. Note that the distribution of $\underline{\Theta}_1$ will in general be different from that of $\underline{\Theta}_2$. It is convenient to denote the combined two group IRT representation by

$$\{d, \underline{\Theta}_g, F_g(\underline{\theta}), \{P_{ig}(\underline{\theta}): \quad i = 1, \ldots N\}: \quad g = 1, 2\}. \tag{4-4}$$

The IRT representation (4-4) will be assumed throughout the remainder of Section 4 (with (3-4), Assumption 3.1, and Assumption 3.2 assumed to hold within each group of course). Implicit in (4-4) is the assumption that the test measures the *same* psychometrically-defined ability construct $\underline{\theta}$ in both groups.

Two basic assumptions additional to Assumption 3.1 and 3.2 about the IRT representation (4-4) are necessary: (1) common multidimensional IRFs in for each of the two groups in the representation (4-4) (i.e., IRF *invariance* across group) and (2) the capability of the test to measure (possibly with contamination) the intended-to-be-measured ability (*target ability*):

Assumption 4.1. *In the assumed IRT representation (4-4) assume IRF group invariance, that is*

$$P_{i1}(\underline{\theta}) = P_{i2}(\underline{\theta}) \equiv P_i(\underline{\theta}) \tag{4-5}$$

for all $\underline{\theta}$. □

This first additional assumption states that the usual IRT item parameter invariance assumed in unidimensional IRT modeling is assumed to hold for our multidimensional IRT model, where $\underline{\theta}$ includes *all* the abilities influencing test performance (hence the assumption of IRF group invariance *is* appropriate in this context). Such invariance does not necessarily hold for any subset of the

components of $\underline{\theta}$, in particular not for the target ability alone. Indeed if invariance with respect to target ability held for all items it is intuitively clear there could be no bias. For example, in the usual definition of item bias (Definition 4.1) invariance is not assumed for the biased item. (4-5) is assumed throughout the rest of the paper.

We now define target ability.

Definition 4.2. *Target ability is the unidimensional latent ability the test intends to measure. The target ability component is denoted by θ, and the target ability random variable for Group g is denoted Θ_g.*

Remark. Θ_g is not to be confused with $\underline{\Theta}_g$ as defined in (4-4).

If a discussion of test bias is appropriate in a test administration, then it must be the case that the test is designed so that it is in fact measuring θ, as well as possibly some nuisance components inadvertently. We thus informally make the second additional assumption that all items of the test in fact do measure target ability θ and possibly nuisance components $\underline{\eta}$ as well. That is, all IRFs $P_i(\theta, \underline{\eta})$ are assumed strictly increasing in θ throughout the paper. In Shealy (1989), this assumption is formalized and it is then proved that the existence of a representation (4-4) in turn implies the existence of an analogous representation in terms of $(\theta, \underline{\eta})$; that is in terms of target ability and nuisance components. Here we bypass presentation of this formalism and instead assume an IRT representation of the form (4-4) with

$$\underline{\Theta}_g \equiv (\Theta_g, \underline{\eta}_g) \tag{4-6}$$

where Θ_g denotes target ability and $\underline{\eta}_g$ denotes nuisance ability for a randomly chosen group g examinee. That is, the two group IRT representation

$$\{d, (\Theta_g, \underline{\eta}_g), F_g(\theta, \underline{\eta}), \{P_i(\theta, \underline{\eta}), \quad i = 1, \ldots, N\} : \quad g = 1, 2\} \tag{4-7}$$

where the P_i's are the group invariant IRFs guaranteed to exist by Assumption 4.1, is assumed throughout the remainder of the paper.

4.3 A Multidimensional Formulation of Test Bias

Iten bias postulates that examinees scaled on a univariate latent θ (as in Definition 4.1) display differing item response probability across group for the biased item. We will take the postulated

ability θ to be the target ability to create an IRT-based definition of test bias.

As in item bias studies, test bias of this sort is an entity studied at the "micro level" of each fixed value of θ; so one may speak of "test bias at θ". Test bias at the "macro level" may be defined to exist if it exists at one or more single θ-values; important aspects of this micro/macro duality are considered in Section 6. The following formulation of test bias is composed of three components:

(a) The *potential for bias*, if it exists, resides within the multidimensional target/nuisance ability distributions in two groups;

(b) potential for bias is *expressed* in items whose responses depend on one or more nuisance determinants; and

(c) the scoring method of the test, to be viewed as an estimate of target ability, transmits expressed item biases into test bias.

4.3.1 Potential for test bias

Before the concept of "potential for test bias" can be developed, it is necessary to introduce conditions postulating stochastic ordering of ability distributions.

Consider a nuisance ability η_g, assumed unidimensional for simplicity of explication, for two groups $g, g = 1, 2$. Either the distribution is the same for both groups or, by definition, there exists some η for which

$$P[\eta_1 > \eta] \neq P[\eta_2 > \eta].$$

Say that, as psychometricians, we believe that Group 1 has "more" of this ability. Likely the most natural way to mathematize this belief is to assume stochastic ordering, that is to assume

$$P[\eta_1 > \eta] > P[\eta_2 > \eta]$$

for *all* η. For η_1 and η_2 that possess densities, the graphical intuition is given in Figure 1. For example, as Figure 1 suggests, the densities might be identical except for translation. Of course, if two groups differ in ability distribution, it does *not* follow logically that one or the other group has "more" ability. For example, a situation where the variances of η_1 and η_2 are not equal can produce

$$P[\eta_1 > \eta] < P[\eta_2 > \eta] \quad \text{for } \eta > 0$$

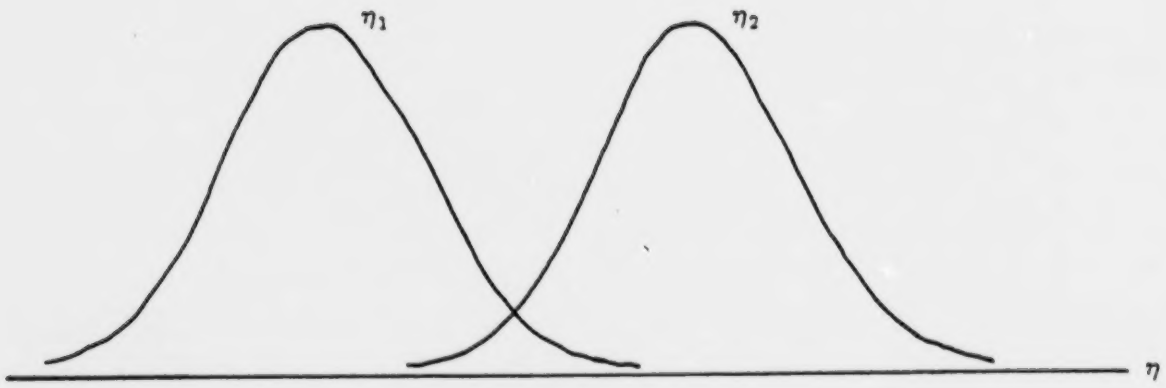

Figure 1:

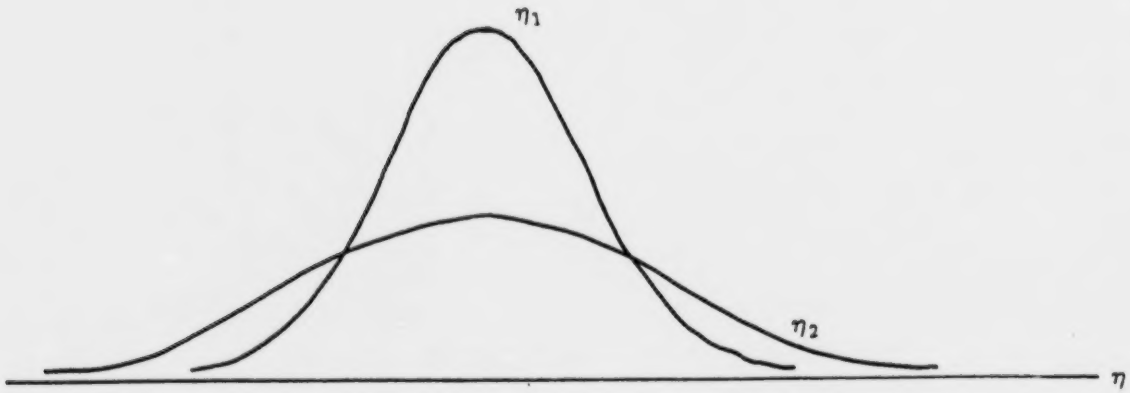

Figure 2:

and

$$P[\eta_1 < \eta] < P[\eta_2 < \eta] \quad \text{for } \eta > 0.$$

In particular, η_1 and η_2 might be symmetrically distributed about 0 with η_2 having the larger variance, as displayed in Figure 2. Nonetheless, for many psychometric applications, it seems plausible to assume stochastic ordering whenever ability distributions are not equal, as we will do below.

The potential for test bias is modeled via one or more determinants that simultaneously cause bias in a collection of items. In particular, this cause is rooted in the conditional distributions of $\underline{\eta}_g \mid \Theta_g = \theta$ (note that $\underline{\eta}_g$ can be multidimensional here). For a fixed θ, we assume stochastic ordering for the distributions of $\underline{\eta}_g \mid \Theta_g = \theta (g = 1, 2)$ when they are not equal:

Assumption 4.2. *Let $(\Theta_g, \underline{\eta}_g)$ be as in (4-7) and fix a target ability value θ. If the conditional distributions $\underline{\eta}_1 | \Theta_1 = \theta$ and $\underline{\eta}_2 \mid \Theta_2 = \theta$ are different, then the assumption is that either*

$$(\underline{\eta}_1 \mid \Theta_1 = \theta) < (\underline{\eta}_2 | \Theta_2 = \theta) \quad \text{or} \quad (\underline{\eta}_1 | \Theta_1 = \theta) > (\underline{\eta}_2 | \Theta_2 = \theta) \tag{4-8}$$

stochastically; i.e., either

$$P[\underline{\eta}_1 > \underline{\eta} \mid \Theta_1 = \theta] < P[\underline{\eta}_2 > \underline{\eta} \mid \Theta_2 = \theta] \tag{4-9}$$

for all $\underline{\eta}$ or

$$P[\underline{\eta}_1 > \underline{\eta} \mid \Theta_1 = \theta] > P[\underline{\eta}_2 > \underline{\eta} \mid \Theta_2 = \theta] \tag{4-10}$$

for all $\underline{\eta}$. □

For example, let θ be mathematical ability and $\underline{\eta} \equiv \eta$ be verbal ability. Then (4-9) says among all examinees of Mathematical Ability θ that, *stochastically*, Group 2 examinees are verbally superior to Group 1 examinees.

With the above preparation, *potential for test bias* can be defined.

Definition 4.3. *Let two groups have ability distributions $(\Theta_1, \underline{\eta}_1)$ and $(\Theta_2, \underline{\eta}_2)$. Potential for test bias exists with respect to nuisance determinant $\underline{\eta}$ at target ability level θ if either (4-9) or (4-10) holds. If (4-9) holds a potential disadvantage exists against Group 1 at target ability θ.* □

Definition 4.3 implies that a potential disadvantage can exist only if there is a nuisance determinant as a component of the latent ability vector.

4.3.2 Expression of test bias potential

In order for test bias to occur, its potential must be *expressed* in one or more items. The concept of expressed bias, detailed in Definition 4.5 below, is similar to the *item bias* concept of Definition 4.1. It is stated in terms of the *marginal IRFs* with respect to target ability:

Definition 4.4. *Refer to (4-7). The marginal IRF*

$$\begin{aligned} T_{ig}(\theta) &= E[P_i(\theta_g, \underline{\eta}_g)|\Theta_g = \theta] \\ &= P[U_i = 1 \mid \Theta_g = \theta] \qquad i = 1, \ldots, N \end{aligned}$$

is called the target marginal IRF for item i, Group g. □

We can now define expressed bias in item i at target ability θ.

Definition 4.5. *Let $\{T_{ig}(\theta) : i = 1, \ldots, N\}$ be Group g's target marginal IRFs for a test with IRT representation (4-7).*

(i) Expressed bias in item i exists at target ability θ if item i's target marginal IRF for Group 1 is not equal to the corresponding target marginal IRF for Group 2 at θ:

$$T_{i1}(\theta) \neq T_{i2}(\theta).$$

(ii) Expressed bias in item i exists if there is some value θ for which expressed bias for item i exists at θ.

Item i is biased against Group 1 at θ if $T_{i1}(\theta) < T_{i2}(\theta)$. □

Definition 4.5 (our multidimensional IRT expressed item bias definition) is equivalent to Definition 4.1 (the usual IRT item bias definition) if

(i) the IRT models represented by (4-2) and (4-7) are both IRT representations of $\{\underline{U}_g : g = 1, 2\}$,

(ii) the ability θ of (4.2) is the target ability θ of Definition 4.2, and

(iii) the group-dependent IRF $P_{ig}(\cdot)$ from (4-2) is taken to be the target marginal IRF $T_{ig}(\cdot)$ from Definition 4.4.

Henceforth in the paper, "item bias" will refer specifically to the expressed item bias of Definition 4.5.

The link between potential for bias and expressed bias for an item is the heart of test bias. The following theorem is fundamental in establishing this link.

Theorem 4.1. *Assume IRT representation (4-7) and fix the number θ. If $P_i(\theta, \underline{\eta})$ is strictly increasing in $\underline{\eta}$ and a potential disadvantage exists against Group 1 at θ then item i is biased against Group 1 in the sense of Definition 4.5.*

Proof. The result is an immediate corollary of Theorem 3.1.

Remark. In a sense, Theorem 4.1 formalizes the obvious; dependence of an item on nuisance determinants with respect to which one group is disadvantaged causes expressed item bias.

4.3.3 Transmission of expressed item biases into test bias

Until now the discussion has focused on a single item; we shall see that a test can consist of many items simultaneously biased by the same nuisance determinant. In this case, items can cohere and act through the prescribed test score to produce substantial bias against a particular group even if individual items display undetectably small amounts of item bias.

This is the final component of our formulation of test bias mentioned at the beginning of this section. We consider the large class of test scores of the form

$$h(\underline{U}) \tag{4-11}$$

where $h(\underline{u})$ is real valued with domain all $\underline{u} \equiv (u_1, \ldots, u_N)$ such that $u_i = 0$ or 1 for $i = 1, \ldots, N$ and $h(\underline{u})$ is coordinate wise non-decreasing in $\underline{u}$. This class contains many of the standard scoring procedures for many standard models; for example, number correct, linear formula scoring of the form $\sum_{i=1}^{N} a_i U_i$, with $a_i \geq 0$, maximum likelihood estimation of ability for certain logistic models with item parameters assumed known, etc. One is surely willing to restrict attention to test scores of the form (4-11), if the test's IRFs are known to be increasing. Following Rosenbaum (1985), test scores of the form (4-11) will be called *non-decreasing item summaries.*

Test bias is defined with respect to a specific test scoring method $h(\underline{u})$.

Definition 4.6. *A test $\underline{U}$ with target ability Θ and test score $h(\underline{U})$ displays test bias against Group 1 at θ if*

$$E[h(\underline{U}_1) \mid \Theta_1 = \theta] < E[h(\underline{U}_2) \mid \Theta_2 = \theta]. \tag{4-12}$$

If

$$E[h(\underline{U}_1) \mid \Theta_1 = \theta] = E[h(\underline{U}_2) \mid \Theta_2 = \theta] \qquad (4-13)$$

then no test bias exists at θ. □

The psychometric interpretation of Definition 4.6 is as follows. The left side of (4-13) is the expected test score for a randomly chosen Group 1 examinee with target ability θ while the right hand side is the same for a randomly chosen Group 2 examinee with target ability θ. In order to assess the appropriateness of Definition 4.6, consider a large number of Group 1 and a large number of Group 2 examinees taking the test, all of target ability θ. Then (4-13) says that the average score of these Group 1 examinees will be approximately the same as that of the Group 2 examinees. Thus, *on average*, neither group is favored in the attempt to estimate target ability using $h(\underline{U}_g)$.

4.3.4 A fundamental relationship

We now elucidate how the three conceptual components of our formulation interact to produce test bias. For ease of interpretation we restrict ourselves to the case of a unidimensional η; however, the following results hold if a vector-valued nuisance determinant $\underline{\eta}$ is assumed.

The basic test bias result is given in Theorem 4.2, namely the precise mechanism by which potential for bias is transmitted into test bias. First a variation of a well-known lemma is needed, which for convenience is specialized to the present setting.

Lemma 4.1. *Let* $f(\eta)$ *be strictly increasing in* η *and let stochastic ordering in the sense of (4-9) hold for each fixed* θ. *Then for each fixed* θ

$$E[f(\eta_1)|\Theta_1 = \theta] < E[f(\eta_2)|\Theta_2 = \theta].$$

Proof. Fix θ and let $F_g(\eta)$ denote the cdf of $f(\eta_g) \mid \Theta_g = \theta$. Assume, for simplicity of argument and without loss of generality, that $F_i(0) = 0$ for $g = 1, 2$. Then

$$E[f(\eta_g) \mid \Theta_g = \theta] = \int_0^\infty x dF_g(x).$$

Integration by parts yields

$$\int_0^\infty x dF_g(x) = \int_0^\infty (1 - F_g(x))dx. \qquad (4-14)$$

But (4-9) and $f(\eta)$ strictly increasing implies that

$$F_1(x) > F_2(x) \text{ for all } x > 0.$$

Using (4-14) and noting that

$$\int_0^\infty (1 - F_1(x))dx < \int_0^\infty (1 - F_2(x))dx,$$

the desired result follows. □

The theory of associated random variables is helpful in establishing the basic test bias result. As defined by Esary, Proschan, and Walkup (1967), a random vector $\underline{X}$ is associated if, and only if, for all nondecreasing $f(\underline{x})$, $g(\underline{x})$, it follows that

$$cov(f(\underline{X}), g(\underline{X})) \geq 0. \tag{4-15}$$

The main result of Esary, Proschan, and Walkup (1967) that we wish to use is that a vector of independent random variables is associated. The basic result can now be stated and proved.

Theorem 4.2. *Assume IRT representation (4-7) with $\underline{\eta} \equiv \eta$ being unidimensional. Fix the number θ and assume the test scoring method of the form (4-11). Suppose for some i that $h(\underline{u})$ is strictly increasing as $u_i = 0$ increases to $u_i = 1$ and that $P_i(\theta, \eta)$ is strictly increasing in η. Assume potential for bias at θ against Group 1, i.e., that (4-9) holds. Then test bias at θ against Group 1 holds.*

Proof. It suffices to prove (4-12). By IRF invariance with respect to (θ, η), it follows for all η and the fixed θ that

$$E[h(\underline{U}_1) \mid \Theta_1 = \theta, \eta_1 = \eta] = E[h(\underline{U}_2) | \Theta_2 = \theta, \eta_2 = \eta] \tag{4-16}$$

Conditioning on $\Theta_g = \theta$, $\eta_g = \eta$ will be denoted by θ, η. Let

$$f(\eta) \equiv E[h(\underline{U}_g) | \theta, \eta],$$

Note that $f(\eta)$ does not depend on g by (4-16), hence let $\underline{U} \equiv \underline{U}_1$ throughout the remainder of the proof. We first show that $f(\eta)$ is strictly increasing in η. Fix $\eta' > \eta$. Then, by local independence

$$q(\underline{u}) \equiv \frac{P[\underline{U} = \underline{u} \mid \theta, \eta']}{P[\underline{U} = \underline{u} \mid \theta, \eta]} = \frac{\prod_{i=1}^N P_i(\theta, \eta')^{u_i}(1 - P_i(\theta, \eta'))^{1-u_i}}{\prod_{i=1}^N P_i(\theta, \eta)^{u_i}(1 - P_i(\theta, \eta))^{1-u_i}}.$$

Thus, $q(\underline{u})$ is strictly increasing as $u_i = 0$ increases to $u_i = 1$ because $P_i(\theta,\eta')/P_i(\theta,\eta) > 1$. Now

$$\begin{aligned} E(h(\underline{U}) \mid \theta,\eta') - E(h(\underline{U})|\theta,\eta) &= \textstyle\sum_{\underline{u}} h(\underline{u})(P[\underline{U}=\underline{u}|\theta,\eta'] - P[\underline{U}=\underline{u} \mid \theta,\eta]) \\ &= \textstyle\sum_{\underline{u}} h(\underline{u})[q(\underline{u})-1]P[\underline{U}=\underline{u} \mid \theta,\eta] \\ &= cov(h(\underline{U}), q(\underline{U})-1|\theta,\eta). \end{aligned} \qquad (4-17)$$

Partition

$$\underline{U} = (\underline{U}\backslash U_i, U_i) \equiv (\underline{U}', U_i)$$

where

$$\underline{U}\backslash U_i \equiv (U_1,\ldots,U_{i-1},U_{i+1},\ldots,U_N).$$

Let E_W and cov_W denote expectation and covariance over the distribution of W, respectively. By a basic identity for covariance, stated here conditional on (θ,η),

$$\begin{aligned} cov(h(\underline{U}), q(\underline{U})-1 \mid \theta,\eta) &= E_{\underline{U}'}\{cov_{U_i}[h(\underline{U}), q(\underline{U})-1|\theta,\eta] \mid \theta,\eta\} \\ &+ cov_{\underline{U}'}\{E_{U_i}(h(\underline{U}) \mid \theta,\eta), E_{U_i}(q(\underline{U})-1 \mid \theta,\eta) \mid \theta,\eta\}. \end{aligned} \qquad (4-18)$$

Both $h(\underline{u})$ and $q(\underline{u}) - 1$ are strictly increasing as $u_i = 0$ increases to $u_i = 1$. Thus, for all possible values of U′,

$$cov_{U_i}[h(\underline{U}), q(\underline{U})-1|\theta,\eta] > 0.$$

Thus, the first term on the right hand side of (4-18) is *strictly* positive. Because of the association of independent random variables and the fact that $\underline{U}'$ given θ,η has independent components, it follows that the second term on the right hand side of (4-18) is nonnegative, using also the fact that

$$E_{U_i}(h(\underline{U}) \mid \theta,\eta) \text{ and } E_{U_i}(q(\underline{U})-1 \mid \theta,\eta)$$

are nondecreasing in $\underline{U}'$. Thus,

$$cov(h(\underline{U}), q(\underline{U})-1 \mid \theta,\eta) > 0.$$

But, recalling (4-17),

$$E(h(\underline{U}) \mid \theta,\eta') - E(h(\underline{U}) \mid \theta,\eta) > 0;$$

that is, $f(\eta)$ is strictly increasing in η, as claimed. Then, applying Lemma 4.1 and (4-9) to $f(\eta)$ above, it follows that (4-12) holds, as required. □

Remarks.

(i) It is important to reemphasize that Theorem 4.2 holds if a vector valued nuisance parameter $\underline{\eta}$ is assumed, provided (4-9), the potential for bias at θ, holds for $\underline{\eta}$. That is, the nuisance determinants $\eta_1, \ldots, \eta_{d-1}$ must each create bias in the *same* direction, say against Group 1.

(ii) Stripped of its test bias context and stated as a general theorem about IRT models, a minor variant of Theorem 4.2 with $<$ replaced by $\leq$ at appropriate places is due to Rosenbaum (1985). For our purposes, strict inequality is needed. The proof of Rosenbaum's result is similar to our proof.

A final interesting relationship to note is that the presence of test bias implies the potential for test bias:

Theorem 4.3. *Suppose that test bias against Group 1 holds at θ in the sense of (4-12). Then the potential for bias against Group 1 at θ exists in the sense that (4-9) holds.*

Proof. Recall (4-16), replacing η by $\underline{\eta}$ there. Thus for $g = 1, 2$, it holds that

$$E[h(\underline{U}_g) \mid \Theta_g = \theta] = \int E[h(\underline{U}_1) \mid \Theta_1 = \theta, \underline{\eta}_1 = \underline{\eta}] dF_g(\underline{\eta} \mid \theta) \qquad (4-19)$$

where $F_g(\underline{\eta} \mid \theta)$ is the cdf of $\underline{\eta}_g \mid \Theta_g = \theta$. Suppose (4-12). Thus, using (4-19) for $g = 1, 2$ it follows that

$$\int E[h(\underline{U}_1) \mid \Theta_1 = \theta, \underline{\eta}_1 = \underline{\eta}] dF_1(\underline{\eta} \mid \theta) < \int E[h(\underline{U}_1) \mid \Theta_1 = \theta, \underline{\eta}_1 = \underline{\eta}] dF_2(\underline{\eta} \mid \theta).$$

But this implies that the distributions of $\underline{\eta}_1 \mid \Theta_1 = \theta$ and $\underline{\eta}_2 \mid \Theta_2 = \theta$ are different. Thus, invoking Assumption 4.3, it follows that (4-9) holds. □

4.4 Item Bias Cancellation

As discussed above, and epitomized by Theorem 4.2, items can combine to amplify bias at the test level. In contrast, items displaying bias can also tend to cancel each other out, thus producing little or no bias at the test level. This becomes possible only when the nuisance determinant $\underline{\eta}$ is multidimensional with some of its components displaying potential for bias against Group 1 and others displaying potential for bias against Group 2. The amount of expressed test bias will be a result of the amount of cancellation at the test level and will be dependent on the particular

test score $h(\underline{u})$ used. The theme of cancellation has been presented by Humphreys (1986) and Roznowski (1987) in the non-IRT classical predictive validity context.

The following example illustrates how cancellation can function to produce negligible test bias.

Example 4.1. A test of length N (N an even number for convenience) intended to measure calculation skills has IRT representation

$$\{\{d = 3, (\Theta_g, \eta_{1g}, \eta_{2g}), F_g(\theta, \eta_1, \eta_2), \{P_i(\theta, \eta_1, \eta_2) : \quad i = 1, \ldots, N\}\}, \quad g = 1, 2\}$$

where θ = mathematics skills, η_1 = physics knowledge, and η_2 = reading knowledge. Let S_1 be a subtest with IRFs

$$\{P_i(\theta, \eta_1) : i = 1, \ldots, \frac{N}{2}\}$$

(subtest containing problems with a mathematical physics flavor) strictly increasing in η_1 for every θ and S_2 be a subtest with IRFs

$$\{P_i(\theta, \eta_2) : i = \frac{N}{2} + 1, \ldots, N\}$$

(subtest containing mathematical "word problems") strictly increasing in η_2 for every θ. Suppose that the ith physics IRF is identical to the ith word problem IRF, which is the $(\frac{N}{2} + i)$th item.

Now, condition on a particular mathematics ability θ, and assume for examinees of ability θ that Group 2 has greater knowledge of physics and Group 1 has greater reading skill. So $\eta_{12} \mid \theta > \eta_{11} \mid \theta$ stochastically and $\eta_{21} \mid \theta > \eta_{22} \mid \theta$ stochastically for each choice of θ. Say that this holds for each choice of θ. Furthermore suppose that as distributions, $\eta_{12} \mid \theta = \eta_{21} \mid \theta$ and $\eta_{11} \mid \theta = \eta_{22} \mid \theta$ for all θ. Then by Theorem 4.2, if subtest S_1 were the entire test, it would exhibit test bias against Group 1 at θ for every θ. By contrast if S_2 were the entire test, it would exhibit test bias against Group 2 at θ for every θ. But, for a large class of test scores–those giving approximately equal weight to the S_1 items and to the S_2 items–almost total cancellation of the item biases could occur thus producing an unbiased test. That is, for such a test scoring method $h(u)$,

$$E[h(\underline{U}_1) \mid \Theta_1 = \theta] \doteq E[h(\underline{U}_2) \mid \Theta_2 = \theta]$$

for every θ. Indeed if $h(\underline{u})$ is number correct, then exact equality and hence total cancellation results.

Remark. Note that the concept of test bias compares groups, not individuals. For a particular examinee, a test might be biased against her, even though the test is not biased against Group 1 of which she is a member. This important aspect of bias is an unfortunate consequence of the multidimensional nature of items in most tests. Moreover, it is also a consequence of the unfortunate (and perhaps economically unavoidable) fact that only statistical (i.e., group-level) bias analysis is done, as opposed to individual case-by-case analysis. The above discussed phenomenon of cancellation could possibly alleviate the impact at the individual examinee level (as well, as just discussed, as at group level).

It is worthwhile to develop item bias cancellation in a formal manner.

Definition 4.7. *Item bias cancellation at θ is said to occur if the test consists both of items biased against Group 1 at θ and items biased against Group 2 at θ.*

Remark. It is theoretically possible that cancellation could occur within an item if the item depends on at least two nuisance dimensions, as contrasted with the between item cancellation of Definition 4.7. This source of cancellation, which seems less likely to occur in practice, is not considered in this paper.

Intuitively, the presense of expressed item bias and no cancellation implies test bias. This is the content of Theorem 4.4.

Theorem 4.4. *Assume that at least one item displays expressed item bias at θ in the sense of Definition 4.5, and assume that no item bias cancellation occurs at θ. Then test bias occurs at θ in all non-decreasing item summary test scores $h(\underline{u})$ (see (4-11)) provided $h(\underline{u})$ is strictly increasing in at least one coordinate corresponding to one of the biased items.*

Proof. At the item level, each item is either biased only against one group (Group 1, say) or displays no expressed bias by the assumption of no cancellation. Thus, for all i,

$$P[U_{i1} = 1 \mid \Theta_1 = \theta] \leq P[U_{i2} = 1 \mid \Theta_2 = \theta] \qquad (4-20)$$

with strict inequality for at least one i. Now, by item invariance, for all i,

$$P[U_{i1} = 1 \mid \Theta_1 = \theta, \underline{\eta}_1 = \underline{\eta}] = P[U_{i2} = 1 \mid \Theta_2 = \theta, \underline{\eta}_2 = \underline{\eta}] \equiv P_i(\theta, \underline{\eta}).$$

Recall Assumption 4.2. Note that, denoting the cdf of $(\underline{\eta}_g \mid \Theta_g = \theta)$ by $F_g(\underline{\eta} \mid \theta)$,

$$P[U_{ig} = 1 | \Theta_1 = \theta] = \int P_i(\theta, \underline{\eta}) dF_g(\underline{\eta} \mid \theta)$$

where the integrand does not depend on g. It follows from Assumption 4.2 that strict inequality in (4-20) for some i implies that $(\underline{\eta}_1 \mid \Theta_1 = \theta) < (\underline{\eta}_2 | \Theta_2 = \theta)$ stochastically. Thus using the monotone condition for $h(\underline{u})$ the conclusion follows from Theorem 4.2, noting the remark following the proof of Theorem 4.2 concerning multiple nuisance determinants. □

It is interesting to note, as Theorem 4.5 now states, that when there is no item bias cancellation that test bias for number correct is equivalent to test bias for all nondecreasing item summary test scores with strict increase for at least one coordinate of $\underline{u}$.

Theorem 4.5. *(a) If test bias at θ occurs for the test score number correct $(\sum_{i=1}^{n} u_i)$ and there is no item bias cancellation at θ, then test bias occurs at θ for every nondecreasing item summary test score $h(\underline{u})$ for which $h(\underline{u})$ is strictly increasing in at least one coordinate of $\underline{u}$. (b) If test bias at θ holds for some nondecreasing item summary test score $h(\underline{u})$ and there is no item bias cancellation at θ, then test bias at θ hold for $h(\underline{u}) = \sum_{i=1}^{n} u_i$.*

Proof. Note that

$$E[\sum_{i=1}^{N} U_{ig} \mid \Theta_g = \theta] = \sum_{i=1}^{n} \int P_i(\theta, \underline{\eta}) dF_g(\underline{\eta} \mid \theta).$$

Then, obvious and minor modifications in the proof of Theorem 4.4 suffice to prove both (a) and (b). Details are omitted. □

Intuitively, no test bias and no cancellation implies that none of the items display bias. This is the content of Theorem 4.6.

Theorem 4.6. *Assume that no test bias exists at θ with respect to score $h(\underline{u})$. Assume no item bias cancellation at θ in the sense of Definition 4.7. In addition, assume that there exists at least one i such that both $P_i(\theta, \underline{\eta})$ is strictly increasing in $\underline{\eta}$ and $h(\underline{u})$ is strictly increasing as $u_i = 0$ increases to $u_i = 1$. Then there is no potential for test bias and (hence) none of the items display item bias.*

Proof. By assumption of no test bias at θ,

$$E[h(\underline{U}_1) \mid \Theta_1 = \theta] = E[h(\underline{U}_2) \mid \Theta_2 = \theta]. \qquad (4-21)$$

By the strict increasing assumption for $P_i(\theta, \underline{\eta})$ and $h(\underline{u})$, it follows that $E[h(\underline{U}_g) \mid \theta, \underline{\eta}]$ is strictly increasing in $\underline{\eta}$. Recall (4-19). If either (4-9) or (4-10) were to hold, it would thus be impossible for (4-21) to hold. Thus by regularity Assumption 4.2, it follows that $(\underline{\eta}_1 \mid \Theta_1 = \theta) = (\underline{\eta}_2 \mid \Theta_2 = \theta)$ stochastically; i.e., there is no potential for test bias. Referring to Theorem 4.1, we see that none of the items display item bias. □

Remarks.

(i) Assuming a scoring method really dependent on all items and that at least one of the items actually depends on $\underline{\eta}$, Theorem 4.6 implies that if there is potential for bias, then either test bias results or item bias cancellation results (and possibly both result simultaneously).

(ii) Theorem 4.2 and 4.6 can be together interpreted as stating a set of conditions under which the potential for test bias is equivalent to test bias.

4.5 Valid Subtest

Recall the informal definition of a valid subtest from Section 2. As mentioned therein, the reason for requiring a valid subtest to exist is that it is statistically impossible to detect test bias using only data from an ability test *unless* there exists an internal criterion measuring only the target ability; i.e., a valid subtest. Here we formally define the validity of a subtest. Let θ denote the target ability. Recall from Section 4.2 that all IRFs are assumed strictly increasing in θ.

Definition 4.8. *Let $\underline{U}$ be a test response with IRT representation (3-7), let $\underline{\theta} = (\theta, \underline{\eta})$, and let S be a subset of the items $1, \ldots, N$. S is a valid subtest if the IRFs of all items in S depend only on θ; i.e., $P_i(\theta, \underline{\eta}) = P_i(\theta)$ for each i in S.*

Remarks.

(i) From a practical viewpoint one wants S to consist of as many of the items of the test as possible; the statistical power of detecting test bias increases as the proportion of valid items does.

(ii) Consider a specified nondecreasing item summary scoring method $h(\underline{u})$ for a test response $\underline{U}$ (recall (4-11). Suitably restrict this scoring method to a subtest response $\underline{U}'$, denoting it by $h'(\underline{u}')$. For example, if $h(\underline{u}) = \Sigma_{i=1}^{N} u_i/N$, then $h'(\underline{u}') = \Sigma' u_i/N'$ is the obvious "restriction", where N' is the cardinality of $\underline{U}'$ and Σ' denotes summation over the components of $\underline{U}'$. A plausible alternative definition of subtest validity consistent with this paper's emphasis on the expression of bias at the test level expressed through the test score would be to require of $h'(\underline{u}')$ that for all θ, given

$$E[h'(\underline{U}') \mid (\Theta, \underline{\eta}) = (\theta, \underline{\eta})]$$

depends only on θ and not on $\underline{\eta}$. This assertion is equivalent to asserting for all $(\theta, \underline{\eta})$ that

$$E[h'(\underline{U}') \mid (\Theta, \underline{\eta}) = (\theta, \underline{\eta})] = E[h'(\underline{U}') \mid \Theta = \theta]. \qquad (4-22)$$

(4-22) is appealing as a possible definition of subtest validity because it functions in an aggregate way at the test level based on the specified test scoring method "restricted" to the subtest. Evoking the usual empirical interpretation of expectation, (4-22) says that repeated sampling of examinees from ability groups, both with the same value of θ but with any choice of two different values of $\underline{\eta}$ produces on average approximately the same value of $h'(\underline{U}')$, as one would wish a "valid subtest" to do.

Fortunately, however, this alternate and appealing definition is actually *equivalent* to our Definition 4.8, under the natural and mild regularity condition that $h'(\underline{u}')$ be strictly increasing as $u_i = 0$ is increased to $u_i = 1$ for each component u_i of $\underline{u}'$; that is that $h'(\underline{u}')$ must really depend on each of the valid subtest item responses. This assertion follows from a modification of the proof of Theorem 4.2. Thus our definition of subtest validity can be thought of as operating either at the item level (Definition 4.8) or at the test level ((4-22)).

(iii) Assume a two group representation (4-3). It is perhaps interesting to note it is possible for all θ that

$$E[h'(\underline{U}_1') \mid \Theta_1 = \theta] = E[h'(\underline{U}_2') \mid \Theta_2 = \theta] \qquad (4-23)$$

and yet subtest validity not hold. Note here that (4-22), equivalent to subtest validity, implies (4-23); however, (4-23) should *not* be used as a definition of subtest validity. As an extreme example demonstrating this claim, each item of S could be measuring $\underline{\eta}$ alone with

$\underline{\eta}_g$ independent of Θ_g for $g = 1$ and $g = 2$ and $\underline{\eta}_g$ having the same distribution for $g = 1, 2$. Subtest validity obviously does not hold here because the supposed-to-be valid items may be heavily influenced by $\underline{\eta}$; however,

$$\begin{aligned} E[h'(\underline{U}_1') \mid \Theta_1 = \theta] &= E[E\{h'(\underline{U}_1') \mid \Theta_1 = \theta, \underline{\eta}_1\} \mid \Theta_1 = \theta] \\ &= E[E\{h'(\underline{U}_1') \mid \underline{\eta}_1\} \mid \Theta_1 = \theta] \\ &= Eh'(\underline{U}_1') \\ &= Eh'(\underline{U}_2') = \ldots \\ &= E[h'(\underline{U}_2') \mid \Theta_2 = \theta]; \end{aligned} \tag{4-24}$$

so (4-23) does hold here. The point we have just shown is that the absence of test bias (i.e., that (4-23) holds) does *not* imply test invalidity (i.e., that (4-22) fails). Related to this fact, note that test validity for the entire test in the sense that (4-22) holds for all $(\theta, \underline{\eta})$ for some scoring method $h(\underline{u})$ that is increasing in every component u_i of $\underline{u}$ does imply for every θ that no test bias exists. This follows trivially from the fact that test validity for the entire test means that every item depends only on θ.

5 Test Bias: The Long Test Case

The theory of test bias presented in Section 4 shows that if there is at least one nuisance dimension then test bias may be present. It is well known that purely unidimensional tests are rare among typical aptitude and achievement tests (see Ansley and Forsyth (1985), Humphreys (1984), Reckase, Carlson, Ackerman, and Spray (1986), and Yen (1984), among others). The position is summarized well in Humphreys (1984):

> The related problems of dimensionality and bias of items are being approached in an arbitrary and oversimplified fashion. It should be obvious that unidimensionality can only be approximated. ... The large amount of unique variance in items is not random error, although it can be called error from the point of view of the attribute that one is attempting to measure. ... We start with the assumption that responses to items have many causes or determinants.

How does the empirical reality of multiple determinants on a test interact with our multidimensional model of test bias? There are two cases to consider: either the test is "long" or it is "short". By "long" it is meant that the number of items is large enough that asymptotic probabilistic ar-

guments provide a useful approximation to the actual test operating characteristics. For example, for many purposes a test of 40 items can be classified as "long".

In the case of a short test, several of the results in Section 4 are important: First, even if nuisance determinants are present in the items and influence examinee performance, the potential for bias against a group must exist in order for test bias to be possible. Second, if the amount of expressed bias at the item level is sufficiently small, then the amount of bias possible at the test level is bounded above. However, if little or no cancellation occurs, small amounts of bias at the item level can produce a substantial amount of test bias. Indeed, one can imagine a detrimental amount of test bias, but with statistical testing for individual item bias being unable to detect any bias at the item level. Third, the amount of test bias is dependent upon the scoring method, the scoring method being the link between item and test bias. It is possible that some scoring methods might be more robust against the detrimental influence of item bias than others. Fourth, recalling Example 4.1 and the material on item bias cancellation, it is quite possible to minimize, with the help of an aptly chosen scoring method, the amount of test bias by having different biasing influences cancelling each other out. For example, (again recall Example 4.1) if approximately equal numbers of items express approximately equal amounts of bias, respectively against and in favor of Group 1, then provided the scoring method gives approximately equal weight to the two classes of items, little or no test bias should occur. Intuitively, it seems likely that having many minor dimensions in addition to θ might increase the propensity for cancellation and actually result in *less* test bias. However, in spite of certain encouraging aspects of the above remarks, it is *surely* the fact, because of the intrinsic multidimensional nature of ability tests, that serious amounts of test bias are likely when tests are short.

We now turn the discussion to the development of a "long" test scenario. In the study of test bias in a long test, the theory of essential unidimensionality of a test, as developed by Stout (1987, 1989) and refined by Junker (1989a, b) turns out to be useful. First we summarize the relevant concepts of this theory.

A "long" test response $\underline{U}_N$ is conceptualized as being the initial *observed* segment of a potentially *observable* infinite item pool $\{U_i, i \geq 1\}$. It is assumed that whatever process has been used to construct the first N items of the pool (i.e., the observed test $\underline{U}_N$) could have been continued in the same manner to produce $\{U_i, i \geq 1\}$. With this understanding, in order to do asymptotic statistical

theory and for foundational purposes, we study $\{U_i, i \geq 1\}$ instead of $\underline{U}_N = \{U_i, 1 \leq i \leq N\}$, conceptualizing the item pool $\{U_i, i \geq 1\}$ as the "test". A test $\{U_i, i \geq 1\}$ is defined to be essentially unidimensional ($d_E = 1$) if it has an IRT representation with monotone IRFs but instead of requiring local independence (Assumption 3.1), the weaker assumption is required that

$$\frac{\sum_{1 \leq i < j \leq N} |cov(U_i, U_j | \Theta = \theta)|}{\binom{N}{2}} \to 0 \qquad (5-1)$$

as $N \to \infty$ for every θ. (The requirement of monotonicity can be weakened somewhat when modeling items where non-monotonicity is suspected, but we omit discussion here (see Stout, 1989; Junker, 1989b). When $d_E = 1$, it is shown that the latent ability is unique in the sense that any other $d_E = 1$ IRT representation has a latent trait that is a *monotone rescaling* of θ. (E.g., a mathematics test cannot be a test of geography for the reason that there exists no such rescaling.)

We now must specify a class of scoring methods for the sequence of long tests $\{\underline{U}_N, N \geq 1\}$. It is convenient to consider a large class of such scoring methods, but less extensive than the non-decreasing item summaries (4-11). Recall from mathematical analysis that a collection of functions $\{k_N(x)\}$ is *equicontinuous* if for every $\epsilon > 0$ there exists $\delta > 0$ such that

$$|k_N(x) - k_N(y)| < \epsilon$$

for all N and all x, y for which $|x - y| < \delta$. Note that the assumed continuity is uniform both in the argument *and* in the choice of function.

Definition 5.1. *$\{k_N(\sum_{i=1}^N a_{Ni} U_i)\}$ is called an equicontinuous balanced scoring method provided*

(a) $k_N(x)$ is defined on $[0,1]$, is non-decreasing, and satisfies

$$-\infty < \inf_N k_N(0) \leq \sup_N k_N(0) < \inf_N k_N(1) \leq \sup_N k_N(1) < \infty. \qquad (5-2)$$

(b) $\{k_N(x)\}$ is equicontinuous, and

(c) $\{a_{Ni} : 1 \leq i \leq N, N \geq 1\}$ satisfies $0 \leq a_{Ni} \leq C/N$ for some $C > 0$ and for all i, N and $\sum_{i=1}^N a_{Ni} = 1$ for all N.

Remarks.

(i) (5-2) and (c) merely guarantee that the "empirical" scale established by $k_N(\sum_{i=1}^N a_{Ni}U_i)$ does not shrink to 0 or stretch to ∞ as N varies. For example, if $k_N(1) - k_N(0) \to 0$ as $N \to \infty$, then $k_N(\sum_{i=1}^N a_{Ni}U_i)$ for large N is uninteresting.

(ii) The $a_{Ni} \leq C/N$ guarantees that no single item dominates the score; i.e., the scoring is "balanced".

(iii) A remark on notation is appropriate. An arbitrary scoring method $h_N(\underline{U}_N)$ assigns a score to each test response $\underline{U}_N$ and hence $h_N(\cdot)$ is a function with an N-dimensional domain (such a score occurs in (4-11)). By contrast, an equicontinuous balanced scoring method $k_N(\sum_{i=1}^N a_{Ni}U_i)$ assigns a score to each linear combination $\sum_{i=1}^N a_{Ni}U_i$ for each N and hence $k_N(\cdot)$ is a function with a unidimensional domain.

A fundamental result of "long" test theory is that of a test $\{U_i, i \geq 1\}$ is essentially unidimensional, consistent estimation of θ is possible in the sense that for *any* equicontinuous balanced scoring method, given $\Theta_g = \theta$,

$$k_N\left(\sum_{i=1}^N a_{Ni}U_{ig}\right) - k_N\left(\sum_{i=1}^N a_{Ni}T_i(\theta)\right) \to 0 \qquad (5-3)$$

in probability as $N \to \infty$, for $g = 1, 2$ (established by a minor modification of the proof of Theorem 3.2 in Stout (1989)). That is, θ is estimated with total accuracy in the limit, using the latent scale

$$k_N\left(\sum_{i=1}^N a_{Ni}T_i(\theta)\right).$$

Here $T_i(\theta)$ denotes the marginal item response function defined by $T_i(\theta) = E[P_i(\underline{\Theta})|\Theta = \theta]$. Expectation is over both groups here; that is, Θ is the target ability of a randomly chosen examinee from the pooled group resulting from combining the two groups. An important special case is that when $d_E = 1$, given $\Theta_g = \theta$,

$$\sum_{i=1}^N U_{ig}/N - \sum_{i=1}^N P_i(\theta)/N \to 0$$

in probability as $N \to \infty$, for $g = 1, 2$.

Armed with the above concepts, a "long-test" definition of test bias is now given. The intuitive idea is that if the test scoring method being used measures target ability equally well in both groups

as measured by the convergence in probability behavior as $N \to \infty$, then no test bias exists. Let

$$\underline{U}_g = (U_{1g}, \ldots U_{Ng}, U_{N+1,g}, \ldots)$$

denote the infinite item pool for Group g and let

$$\underline{U}_{Ng} = (U_{1g}, \ldots, U_{Ng})$$

denote the finite *observed* segment of the item pool for g. To study long-test test bias, we make the assumption that $\underline{U}_g$ has a two group representation of the form (4-4) with (3-4), Assumptions 3.1 and 3.2 holding within each group and with Assumption 4.1 holding. It then follows from the ordinary weak law of large numbers in probability theory for any equicontinuous balanced test scoring method that, given $\underline{\Theta}_1 = \underline{\theta}$ and $\underline{\Theta}_2 = \underline{\theta}$,

$$\begin{aligned} &k_N\left(\textstyle\sum_{i=1}^N a_{Ni}U_{i1}\right) - k_N\left(\textstyle\sum_{i=1}^N a_{Ni}P_i(\underline{\theta})\right) \to 0 \\ &\text{and} \\ &k_N\left(\textstyle\sum_{i=1}^N a_{Ni}U_{i2}\right) - k_N\left(\textstyle\sum_{i=1}^N a_{Ni}P_i(\underline{\theta})\right) \to 0 \end{aligned} \qquad (5-4)$$

in probability as $N \to \infty$. Here $\underline{\theta} = (\theta, \underline{\eta})$ where θ is the target ability and $\underline{\eta}$ is the nuisance determinant. Of course, in order to be able to assume local independence for the representation (4-4) and have good model fit the dimension d of $\underline{\eta}$ may need to be quite large. It is easy to show (5-4) also holds for an d_E essential dimensional representation of the form (4-4), with d_E possibly much smaller than d.

Because $k_N(\sum_{i=1}^N a_{Ni}U_{i1})$ and $k_N(\sum_{i=1}^N a_{Ni}U_{i2})$ have the same limit behavior in probability (hence $(\theta, \underline{\eta})$ is measured equally well in both groups), (5-4) seems to suggest that no test bias in a long-test sense is possible. However, (5-4) is not the same as group-equivalent measurement of target ability θ alone. As in the finite test length case of Section 4, the source of bias is that the conditional distributions of $(\underline{\eta}_1|\Theta_1 = \theta)$ and $(\underline{\eta}_2|\Theta_2 = \theta)$ differ, thereby leading to superior limiting test scores for one group versus another given $\Theta_1 = \theta$, $\Theta_2 = \theta$. An example should clarify this claim.

Example 5.1. Consider examinee subpopulations from the two groups defined by $\Theta_1 = \theta$ and $\Theta_2 = \theta$, respectively, i.e., both subpopulations have the same target ability. Suppose that there is a single nuisance determinant and that

$$\begin{aligned} P[\eta_1 = 1|\Theta_1 = \theta] &= \tfrac{1}{4} \qquad P[\eta_2 = 1|\Theta_2 = \theta] = \tfrac{3}{4} \\ P[\eta_1 = 0|\Theta_1 = \theta] &= \tfrac{3}{4} \qquad P[\eta_2 = 0|\Theta_2 = \theta] = \tfrac{1}{4} \end{aligned} \qquad (5-5)$$

Clearly this is a case of potential for bias against Group 1 at θ. Suppose $k_N(x) = x$ for all N and $a_{Ni} = 1/N$ for all i and N:

$$k_N\left(\sum a_{Ni}u_i\right) = \sum_{i=1}^{N} u_i/N.$$

Suppose local independence with respect to (θ, η) with

$$P_i(\theta, 1) = \tfrac{2}{3}, P_i(\theta, 0) = \tfrac{1}{3}$$

for all i. Then, (5-4) specializes to

$$\frac{\sum_{i=1}^{N} U_{i1}}{N} \to \tfrac{2}{3}, \quad \frac{\sum_{i=1}^{N} U_{i2}}{N} \to \tfrac{2}{3}$$

given $\Theta_1 = \theta$, $\eta_1 = 1$ and $\Theta_2 = \theta$, $\eta_2 = 1$, respectively, in probability as $N \to \infty$. Also

$$\frac{\sum_{i=1}^{N} U_{i1}}{N} \to \tfrac{1}{3}, \quad \frac{\sum_{i=1}^{N} U_{i2}}{N} \to \tfrac{1}{3}$$

given $\Theta_1 = \theta$, $\eta_1 = 0$ and $\Theta_2 = \theta$, $\eta_2 = 0$, respectively, in probability as $N \to \infty$. But, conditioning on $\Theta_1 = \theta$ and $\Theta_2 = \theta$, it follows using (5-5) that

$$\begin{array}{l} \frac{\sum_{i=1}^{N} U_{i1}}{N} \to \frac{2}{3} \text{ with probability} \frac{1}{4} \text{ and} \\ \frac{\sum_{i=1}^{N} U_{i2}}{N} \to \frac{1}{3} \text{ with probability} \frac{3}{4} \end{array} \qquad (5-6)$$

as $N \to \infty$, as contrasted with

$$\begin{array}{l} \frac{\sum_{i=1}^{N} U_{i2}}{N} \to \frac{2}{3} \text{ with probability} \frac{3}{4} \text{ and} \\ \frac{\sum_{i=1}^{N} U_{i1}}{N} \to \frac{1}{3} \text{ with probability } \frac{1}{4} \end{array} \qquad (5-7)$$

as $N \to \infty$. Clearly Group 2 is favored among examinees of target ability θ. It may be interesting to note that

$$E\left[\frac{\sum_{i=1}^{N} U_{i1}}{N} | \Theta_1 = \theta\right] = \tfrac{3}{4} \cdot \tfrac{1}{3} + \tfrac{1}{4} \cdot \tfrac{2}{3} = \tfrac{5}{12} \qquad (5-8)$$

for all N, while

$$E\left[\frac{\sum_{i=1}^{N} U_{i2}}{N} | \Theta_2 = \theta\right] = \tfrac{1}{4} \cdot \tfrac{1}{3} + \tfrac{3}{4} \cdot \tfrac{2}{3} = \tfrac{7}{12}. \qquad (5-9)$$

Thus, in a trivial manner not dependent on N,

$$\mathrm{Lim}_{N\to\infty} E\left[\frac{\sum_{i=1}^{N} U_{i1}}{N} | \Theta_1 = \theta\right] - E\left[\frac{\sum_{i=1}^{N} U_{i2}}{N} | \Theta_2 = \theta\right] = -\tfrac{1}{6} < 0. \qquad (5-10)$$

□

We will use the idea embodied in (5-10) to define large sample test bias.

Definition 5.2. *Let θ denote target ability.*

(i) There is no long-test test bias at θ with respect to an equicontinuous balanced scoring method

$$k_N\left(\sum_{i=1}^{N} a_{Ni}U_i\right)$$

provided

$$E\left[k_N\left(\sum_{i=1}^{N} a_{Ni}U_{i1}\right)|\Theta_1=\theta\right]-E\left[k_N\left(\sum_{i=1}^{N} a_{Ni}U_{i2}\right)|\Theta_2=\theta\right]\to 0 \qquad (5-11)$$

as $N\to\infty$.

(ii) If for every θ, there is no long-test test bias at θ, then there is no long-test test bias.

(iii) If at θ

$$E\left[k_N\left(\sum_{i=1}^{N} a_{Ni}U_{i1}\right)|\Theta_1=\theta\right]-E\left[k_N\left(\sum_{i=1}^{N} a_{Ni}U_{i2}\right)|\Theta_2=\theta\right]\le C<0 \qquad (5-12)$$

for all sufficiently large N and some C, then long-test test bias exists at θ against Group 1.

□

We first show that if there is no long-test test bias in the empirical sense that among examinees with the same target ability θ neither group is favored in their stochastic test score behavior as $N\to\infty$, then long-test test bias in the sense of Definition 5.2 holds.

Theorem 5.1. *Suppose, given $\Theta_1=\theta$ and $\Theta_2=\theta$ that for an equicontinuous balanced scoring method,*

$$\begin{aligned} &k_N\left(\textstyle\sum_{i=1}^{N} a_{Ni}U_{i1}\right)-c_{N1}(\theta)\to 0 \text{ and}\\ &k_N\left(\textstyle\sum_{i=1}^{N} a_{Ni}U_{i2}\right)-c_{N2}(\theta)\to 0 \end{aligned} \qquad (5-13)$$

in probability for some $c_{N1}(\theta), c_{N2}(\theta)$, as $N\to\infty$. Then (5-11) holds; that is, there is no long-test test bias at θ for the given scoring method.

Remark. Note that it is *not* required that the centering functions $c_{Ng}(\theta)$ have to be the same for $g=1,2$. What *is* required is the existence of a centering function dependent on θ alone and not $\underline{\eta}$ for each g, as contrasted with (5-4). Of course, the case where the centering functions are the same is of special interest and is the main motivation for the theorem, as the remark immediately prior to the statement of the theorem indicates.

Proof. By (5-2), $|k_N(x)| \leq C$ for some $C > 0$. Thus $|c_{Ng}(\theta)| \leq C + D$ for some $D > 0$. By the Lebesgue dominated convergence theorem (see p. 11, Serfling, 1980), using (5-13)

$$E\left[\sum_{i=1}^{N} a_{Ni}U_{ig} - c_{Ng}(\theta)|\Theta_g = \theta\right] \to 0 \qquad (5-14)$$

as $N \to \infty$, for $g = 1, 2$. Now, trivially, the conclusion (5-13) holds given $\Theta_1 = \theta$, $\underline{\eta}_1 = \underline{\eta}$ and $\Theta_2 = \theta$, $\underline{\eta}_2 = \underline{\eta}$ for all $\underline{\eta}$. Thus, subtracting the two results in (5-13),

$$c_{N1}(\theta) - c_{N2}(\theta) \to 0$$

as $N \to \infty$. Let $c_N(\theta) \equiv c_{N1}(\theta)$. It then follows from (5-14) that

$$E\left[\sum_{i=1}^{N} a_{Ni}U_{ig} - c_N(\theta)|\Theta_g = \theta\right] \to 0$$

as $N \to \infty$ for $g = 1, 2$. Subtracting these two limits yields

$$E\left[\sum_{i=1}^{N} a_{Ni}U_{i1}|\Theta_1 = \theta\right] - E\left[\sum_{i=1}^{N} a_{Ni}U_{i2}|\Theta_2 = \theta\right] \to 0$$

as $N \to \infty$, i.e., no long-test test bias exists at θ. □

Remark. We claim that (5-13), (and hence the similar condition (5-3)), is inappropriately strong to use as a definition of lack of long-test bias. To see this, modify Example 5.1 by assuming

$$P[\eta_g = 0|\Theta_g = \theta] = \tfrac{3}{4}, P[\eta_g = 1|\Theta_g = \theta] = \tfrac{1}{4},$$

for $g = 1, 2$. Hence no potential for bias exists. However note that, given $\Theta_1 = \theta$ and $\Theta_2 = \theta$

$$\frac{\sum_{i=1}^{N} U_{Ng}}{N} - \tfrac{2}{3} \to 0 \text{ with probability } \tfrac{1}{4}$$

and

$$\sum_{i=1}^{N} U_{Ng} - \tfrac{1}{3} \to 0 \text{ with probability } \tfrac{3}{4}$$

for both $g = 1$ and $g = 2$. Thus (5-13) is precluded and thus long-test bias *would* be said to exist (even though no potential for bias exists) if (5-13) was made the basis for deciding on the existence of long-test test bias. Note that the above convergence in probability behavior is *identical* for both groups. Intuitively, in this example the estimation of θ by $\sum_{i=1}^{N} U_{Ng}/N$ as $N \to \infty$ is equally *bad* for both groups in the sense that convergence in probability at θ fails to occur in exactly the same manner in both groups. Thus one would *not* wish to claim that test bias is occurring.

The following theorem states that essential unidimensionality is a sufficient condition for ensuring that no long-test test bias exists.

Theorem 5.2. *Suppose $d_E = 1$ for target ability θ in the combined population consisting of Group 1 and Group 2 examinees. Then, with respect to all equicontinuous balanced scoring methods, no long-test test bias exists.*

Proof. Let $\{k_N(\sum_{i=1}^N a_{Ni}U_i)\}$ be an arbitrary equicontinuous balanced scoring method. We need to prove (5-11) for every θ. Fix θ. By work of Stout (1989), $d_E = 1$ implies (5-3) for $g = 1, 2$; i.e., (5-13) holds with $c_{Ng}(\theta) = k_N(\sum a_{Ni}T_i(\theta))$. Thus, by Theorem 5.1, the desired result holds. □

By contrast, if the potential for bias exists at θ, then it follows that there exist balanced scoring methods for which long-test test bias at θ does exist.

Theorem 5.3. *Assume that IRFs are differentiable in η. Let θ denote target ability, η denote the nuisance determinant and assume potential for bias against Group 1 at θ. Assume there exists a balanced scoring method $\{a_{Ni}\}$ (i.e., $k_N(x) = x$ in Definition 5.1) such that at θ,*

$$\frac{d}{d\eta}\sum_{i=1}^{N} a_{Ni}P_i(\theta,\eta) > \epsilon_\eta > 0 \qquad (5-15)$$

for all η and all N. Then long-test test bias exists at θ against Group 1.

Proof. For $\underline{\theta} = (\theta, \eta)$, (5-4) holds given $\Theta_1 = \theta$, $\eta_1 = \eta$; $\Theta_2 = \theta$, $\eta_2 = \eta$. Now, letting $F_g(\eta|\theta)$ denote the cdf of $\eta_g|\Theta_g = \theta$ and using (5-15) and integration by parts

$$\begin{aligned}
&E[\textstyle\sum_{i=1}^N a_{Ni}U_{i1}|\Theta_1 = \theta] - E[\sum_{i=1}^N a_{Ni}U_{i2}|\Theta_2 = \theta] \\
&= \textstyle\int_{-\infty}^{\infty}\{\sum_{i=1}^N a_{Ni}P_i(\theta,\eta)\}d[F_1(\eta|\theta) - F_2(\eta|\theta)] \\
&= -\textstyle\int_{-\infty}^{\infty}\{\frac{d}{d\eta}\sum_{i=1}^N a_{Ni}P_i(\theta,\eta)\}[F_1(\eta|\theta) - F_2(\eta|\theta)]d\eta \\
&\leq \textstyle\int_{-\infty}^{\infty}\epsilon_\eta[F_1(\eta|\theta) - F_2(\eta|\theta)]d\eta \\
&\leq -c(\theta),
\end{aligned}$$

where $c(\theta) > 0$ by the assumption of potential for bias against Group 1. Since this holds for all N, the result is proved by Definition 5.2. □

How is the finite test length definition of test bias (Definition 4.6) related to the long-test test bias definition (Definition 5.1)? The answer is that lack of finite length test bias for all finite length test $\underline{U}_N$ from the item pool $\{U_i, i \geq 1\}$ implies lack of long-test test bias for all equicontinuous balanced test scores.

Theorem 5.4. *Assume an IRT representation for $\{U_i, i \geq 1\}$ of the form (4-4) for $\underline{\theta} = (\theta, \underline{\eta})$. Let $\{k_N(\sum_{i=1}^N a_{Ni}U_i)\}$ be an equicontinuous balanced scoring method. Assume no finite length test*

bias exists; that is, (4-13) holds for all N. Assume regularity Assumption 4.2. Then there is no long-test test bias; that is, (5-11) holds.

Proof. Trivial from examination of (4-13) and (5-11). □

Remark. Of course, long-test test bias holding is less restrictive than finite length test bias holding. Nonetheless it seems an appropriate way to describe biasedness of a test when the test is long.

From the long test perspective, the need to produce a long-test definition of a valid subtest needs to be addressed. Previously in the short test case, our definition of a valid subtest S with response $\underline{U}$ was stated to be equivalent to (4-22) holding for all $(\theta, \underline{\eta})$. Just as the short-test version of no test bias ((4-13)) is modified for the long-test version of no test bias ((5-11)), a similar modification of (4-22) yields an appropriate definition of a valid subtest. We consider only equicontinuous balanced scoring methods for subtests $\underline{U}'_N$ of $\underline{U}_N$. That is, we consider scoring $k'_N(\sum' a_{Ni}U_i)$ where Definition 5.1 holds, for each $k'_N(\sum' a_{Ni}U_i)$ where $\sum'$ denotes summation over the indices of the components of $\underline{U}'_N$.

Definition 5.3. *Let the item pool $\{U_i, i \geq 1\}$ have IRT representation (3-7) with the usual accompanying assumptions. Let $\underline{U}'_N \subset \underline{U}_N$ denote a subtest of $\underline{U}_N$ for each N. Denoting the cardinality of a set A as card (A), assume*

$$\underline{U}'_N \subset \underline{U}'_{N+1}, \quad \frac{\text{card } \underline{U}'_N}{N} \geq C > 0 \tag{5-16}$$

for some C and for all $N \geq N_0$ for some fixed N_0 (N_0 will be small in all applications). Then $\{\underline{U}'_N, N \geq 1\}$ is said to be a collection of valid subtests with respect to a specified equicontinuous balanced scoring method $\{k'_N(\sum' a_{Ni}U_i)\}$ provided there exists a function $c_N(\theta)$ such that for all $\theta, \underline{\eta}$,

$$E[k'_N(\Sigma' a_{Ni}U_i) \mid \Theta, \underline{\eta}) = (\theta, \underline{\eta})] - c_N(\theta) \to 0 \tag{5-17}$$

as $N \to \infty$.

Remark. Recall that short-test bias validity, i.e., (4-22) hold for all $(\theta, \underline{\eta})$, for scoring method $k'_N(\sum' a_{Ni}U_i)$ say, simply means that for θ

$$m(\theta, \underline{\eta}) \equiv E[k'_N(\Sigma' a_{Ni}U_i) \mid (\Theta, \underline{\eta}) = (\theta, \underline{\eta})]$$

depends only on θ and not on $\underline{\eta}$. By contrast the long-test subtest validity just defined by (5-17) weakens this to asserting that $m(\theta, \underline{\eta})$ for all θ is asymptotically not dependent on $\underline{\eta}$ as $N \to \infty$. That is, intuitively, for large fixed N, $m(\theta, \underline{\eta})$ for all θ is approximately constant as $\underline{\eta}$ varies.

As with long-test test bias, the theory of essential unidimensionality is useful in studying long-test subtest validity:

Theorem 5.5. *Assume $d_E = 1$ with latent ability θ being target ability for subtests $\{\underline{U}'_N, N \geq 1\}$ satisfying (5-16). Then (5-17) holds for all equicontinuous balanced scoring methods; i.e., subtest validity holds for all equicontinuous balanced scoring methods.*

Proof. It follows from a minor modification of the proof of Theorem 3.2 in Stout (1989) that for all $(\theta, \underline{\eta})$

$$k'_N(\Sigma' a_{Ni} U_i) - c_N(\theta) \to 0 \qquad (5-18)$$

in probability as $N \to \infty$. But $|k'_N(\sum' a_{Ni} U_i| \leq C$ for some constant $C < \infty$. It is a standard result from the theory of convergence in probability that convergence in probability and the boundedness just stated together imply convergence in expectation. That is, for all $(\theta, \underline{\eta})$,

$$E[k'_N(\Sigma' a_{Ni} U_i) \mid \Theta, \underline{\eta}) = \theta, \underline{\eta}] - c_N(\theta) \to 0$$

as $N \to \infty$. I.e., (5-17) holds. □

Stout (1987) has developed a statistical test for essential unidimensionality. Clearly this could be applied to a subtest to assess whether it can be used as a valid subtest in the case of a "long" test.

6 Test Bias as a Function of Target Ability

Sections 4 and 5 focus on test bias for fixed values of target ability θ. In these sections it was argued that test bias (item bias also) is a phenomenon that expresses itself at each θ. In particular, it is the comparison of the distributions of $(\underline{\eta}_1 | \Theta_1 = \theta)$ and $(\underline{\eta}_2 | \Theta_2 = \theta)$ that dictates whether test bias is possible at θ and if such bias is possible, in which direction (biased in favor of or biased against Group 1) it occurs. Mathematically, without further assumptions, one cannot infer what the character of the bias at $\theta' \neq \theta$ is from the character of the bias at θ. This section develops

the concept of considering test bias aggregated over target ability. We return to the convention of suppressing N in the notation when appropriate; e.g. $\underline{U} \equiv \underline{U}_N$.

Definition 6.1. *Let $h(\underline{U})$ be a test scoring method and $\underline{U}$ be a test response as in (3-1). The expected test bias at θ against Group 1 using test scoring method $h(\underline{U})$ is given by*

$$B(\theta) \equiv E[h(\underline{U}_2)|\Theta_2 = \theta] - E[h(\underline{U}_1)|\Theta_1 = \theta]. \qquad (6-1)$$

□

Remarks.

(i) Note that $B(\theta) > 0$ indicates test bias against Group 1 at θ.

(ii) Several special cases are of interest. If $h(\underline{u}) = \sum_{i=1}^{N} u_i/N$, then $B(\theta)$ is the difference of (marginal) test characteristic curves (average of marginal IRFs):

$$B(\theta) = \frac{\sum_{i=1}^{N} T_{i2}(\theta)}{N} - \frac{\sum_{i=1}^{N} T_{i1}(\theta)}{N}. \qquad (6-2)$$

If $h(\underline{u}) = u_i$, then

$$B(\theta) = T_{i2}(\theta) - T_{i1}(\theta),$$

the amount of item i bias against Group 1 at θ.

Probably the most common pattern in the potential for bias as a function of θ is unidirectional potential for bias:

Definition 6.2. *If potential for bias exists against the same group at every θ then unidirectional potential for bias is said to exist against the group.* □

Another less common, but still important pattern in the potential for bias as a function of θ is that the "direction" of the potential for bias changes from one end of the θ-continuum to the other:

Definition 6.3. *Suppose for some fixed θ_0 that the potential for bias against one group exists for all $\theta < \theta_0$ and the potential for bias exists against the other group for all $\theta > \theta_0$. Then bidirectional potential for bias is said to exist.* □

The verbal analogies example of Section 2 is an obvious practical example of unidirectional potential for bias. For, it seems likely that the potential for test bias against German immigrants will hold *regardless* of the level of verbal analogies ability being conditioned on.

As an example of bidirectional potential for bias, suppose Θ_1 and Θ_2 are both uniformly distributed on the interval $[-1, 1]$. Suppose that in Group 1, Θ_1 and η_1 are statistically independent with η_1 uniformly distributed on $[-1, 1]$. Suppose in Group 2 that $(\eta_2|\Theta_2 = \theta)$ has a uniform distribution on the interval with end points 0 and 2θ. That is, perhaps because of cultural differences, in Group 2 it follows that Θ and η are highly positively correlated while Θ and η are uncorrelated in Group 1. Elementary computation show that if $-1 < \theta < 0$, (4-10) holds, yet if $0 < \theta < 1$, (4-9) holds. That is, potential for bias against Group 2 holds for $\theta < 0$ and potential for bias against Group 1 hold if $\theta > 0$; i.e., bidirectional potential for bias holds.

Test bias (and item bias) can be undirectional or bidirectional.

Definition 6.4. *If test bias (either in the ordering sense of Definition 4.6 or in the long-test sense of Definition 5.2) exists against the same group at every θ, then unidirectional test bias against that group is said to hold.*

Definition 6.5. *If for some θ_0 test bias in the sense of Definition 4.6 holds against one group for all $\theta < \theta_0$ and against the other group for all $\theta > \theta_0$ then bidirectional test bias is said to occur.* □

A long-test version of Definition 6.5 is easy to give but is omitted for simplicity. The following results relate unidirectional potential for bias to unidirectional test bias.

Theorem 6.1. *Suppose test bias exists against Group 1 at some θ in the sense of Definition 4.6, and suppose unidirectional potential for bias. Assume a test scoring method of the form (4-11). Suppose for every θ' that there is some i (possibly dependent on θ') for which $h(\underline{u})$ is strictly increasing as $u_i = 0$ increases to $u_i = 1$ and for which $P_i(\theta', \underline{\eta})$ is strictly increasing in $\underline{\eta}$. Then unidirectional test bias against Group 1 holds.*

Proof. By Theorem 4.3, the potential for bias against Group 1 at θ holds. By assumption of unidirectional potential for bias, the potential for bias against Group 1 thus holds for all θ'. Apply Theorem 4.2 together with the remark (i) following it. □

Theorem 6.2. *Assume IRFs are differentiable in η. Suppose long-test test bias exists against Group 1 at some θ in the sense of Definition 5.2 for a balanced scoring method $\{a_{N_i}\}$ and suppose*

unidirectional potential for test bias. Assume for every θ'

$$\frac{d}{d\eta}\sum_{i=1}^{N} a_{Ni}P_i(\theta',\eta) > \epsilon_\eta > 0$$

for all η *(without loss of generality assumed unidimensional here). Then unidirectional (long-test) test bias against Group 1 holds in the sense of Definition 6.4.*

Proof. Same as that of Theorem 6.1 except Theorem 5.3 is used in place of Theorem 4.2. □

In order to study bidirectional test bias,-attention is restricted to balanced scoring methods. For an arbitrary balanced scoring method $\sum_{i=1}^{N} a_{Ni}U_i$, letting

$$F_g(\eta|\theta) \equiv P[\eta_g \leq \eta|\Theta_g = \theta]$$

and assuming differentiability of IRFs and a unidimensional nuisance determinant, the following formula for $B(\theta)$ of (6-1) obtained by integration by parts is useful

$$B(\theta) = \int_{-\infty}^{\infty} \left\{\sum_{i=1}^{N} a_{Ni}\frac{d}{d\eta}P_i(\theta,\eta)\right\}[F_1(\eta|\theta) - F_2(\eta|\theta)]\,d\eta. \qquad (6-3)$$

Theorem 6.3. *Assume a balanced scoring method with differentiable IRFs. Assume a unidimensional nuisance trait* η*. Assume for each* θ*, there exists some* i *(possibly varying with* θ*) for which*

$$a_{Ni} > 0, \quad \frac{d}{d\eta}P_i(\theta,\eta) > 0 \text{ for all } \eta > 0. \qquad (6-4)$$

Then bidirectional potential for test bias holds if and only if bidirectional test bias holds.

Proof. By Assumption 4.2, for fixed θ either

$$F_1(\eta|\theta) - F_2(\eta|\theta) > 0 \text{ for all } \eta \qquad (6-5)$$

or

$$F_1(\eta|\theta) - F_2(\eta|\theta) < 0 \text{ for all } \eta. \qquad (6-6)$$

Thus, using (6-3), (6-4) and the strict monotonicity of every $P_i(\theta,\eta)$ in η, $B(\theta) > 0$ or $B(\theta) < 0$ *accordingly as* (6-5) or (6-6) holds. Potential for bias at θ means that either (6-5) or (6-6) holds at θ. The desired result follows. □

Assume number correct scoring, which implies (6-2) and hence that test bias is controlled by the (marginal) item response functions with respect to target ability. Graphically, bidirectional

test bias under this scoring method is shown in Figure 3. Note the effect is that the test displays higher discrimination for Group 2 than for Group 1. That is, bidirectional test bias is expressed as differing test discriminations for the two groups. By contrast, under (6-2), unidirectional test bias is shown in Figure 4. Unidirectional test bias is not linked to differing test discriminations across group. Indeed the two test characteristic curves shown in Figure 4 can even be translates of one another; e.g., for some $c > 0$ for given $T_{i2}(\theta) \equiv T_i(\theta)$

$$\sum_{i=1}^{N} T_{i1}(\theta)/N = \sum_{i=1}^{N} T_i(\theta + c)/N$$

for all θ. That is, items could be uniformly more difficult for Group 2 examinees at every θ.

There is a debate about whether from the cognitive perspective, differing discriminations across group is more the essence of bias than differing difficulties across group. Also, some practitioners claim that bidirectional test bias can be important in practice while others discount its importance. It is hoped that Section 6 helps illuminate these issues.

7 Discussion and Summary of Results

The central position of this paper is that bias should be conceptualized, studied, and measured at the test level rather than at the item level. A multidimensional but non-parametric IRT model of test bias is presented and a number of important properties derived. Our theory of test bias includes the often used unidimensional IRT bias approach as a special case.

The model hypothesizes a *target ability* intended to be measured by the test as well as other dimensions called *nuisance determinants*, not intended to be measured. Informally, test bias occurs when the test under consideration is measuring nuisance determinants in addition to the target ability, and moreover the two groups do not possess equal amounts of the nuisance determinants. Our view, an outgrowth of the classical predictive validity viewpoint of bias, is that bias is really something expressed at the test level via the particular test score in use and that bias rests in the across-group differences in the relationship between test scores and criterion. For us the "criterion" is internal to the test and is expressed by a "valid" subtest known to consist of items measuring only target ability. In order to statistically detect test bias, a valid subtest must exist and be identified.

In Section 3, the multidimensional non-parametric IRT model is presented. The notion of the *marginal IRF* with respect to target ability is introduced.

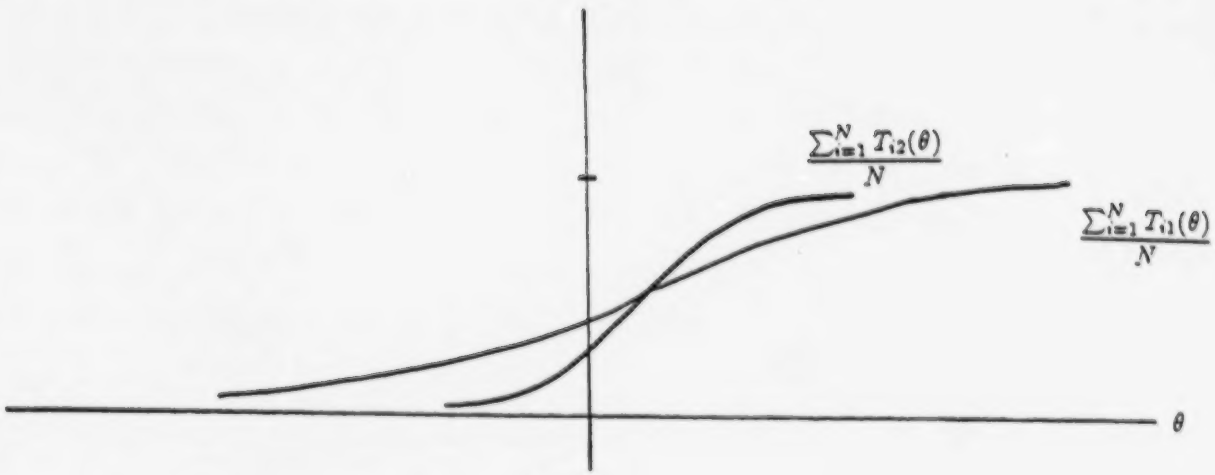

Figure 3:

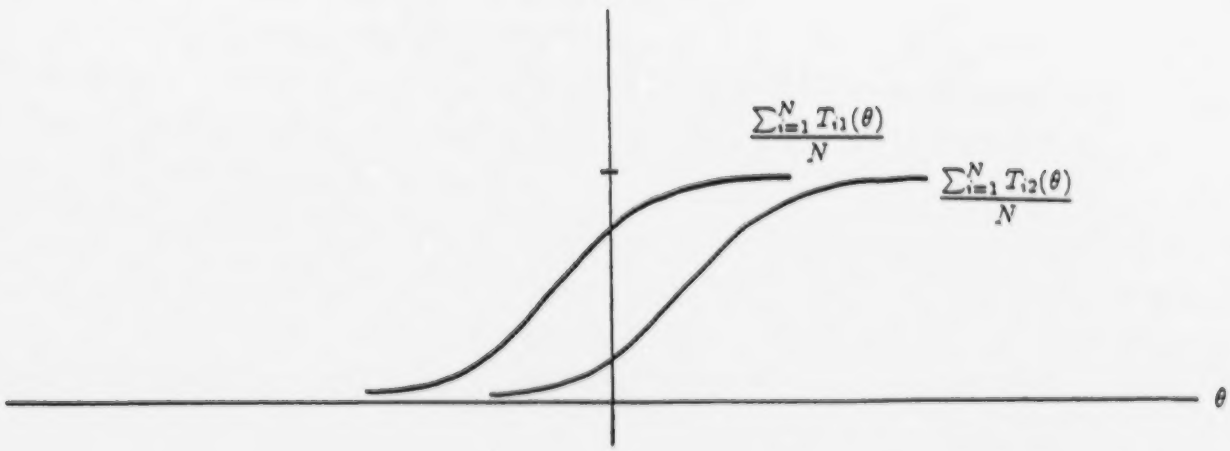

Figure 4:

In Section 4, test bias is carefully defined using the IRT model introduced in Section 3. Test bias originates with the *potential for test bias* at a particular value of θ of target ability existing against a group in the sense of Definition 4.3. This potential for bias against Group 1 gets expressed at the item level if any of the marginal group IRFs satisfy $T_{i1}(\theta) < T_{i2}(\theta)$. The potential for test bias and a strictly increasing IRF in $\underline{\eta}$ implies expressed item bias (Theorem 4.1).

The main focus of this paper is on biased items acting in concert. Three components combine to produce test bias: (a) potential for bias, (b) dependence of the IRFs on $\underline{\eta}$, and (c) the test scoring method, which transmits simultaneous expressed item bias into test bias. Test bias is formally defined in (4-12). It is shown that test bias at θ implies the potential for bias at θ (Theorem 4.3). The central result of Section 4 (Theorem 4.2) shows that potential for bias at θ translates into test bias at θ provided the scoring method depends on at least one item that has a strictly increasing IRF in $\underline{\eta}$ at θ.

The important topic of item bias cancellation is taken up in Section 4.4. Example 4.1 illustrates how cancellation can actually decrease the amount of item bias that gets expressed at the test level. That is, the potential for bias need not be strongly transmitted to the test level because in fact considerable cancellation can occur as the result of multidimensional nuisance determinants. By contrast, small and perhaps undetectable amounts of bias at the item level can be translated into a substantial amount of bias expressed at the test level when no cancellation occurs. Section 4.5 formalizes the notion of a valid subtest, which must exist for text bias to be detected. Shealy and Stout (1990) present a statistical test of test bias, making the question of whether test bias does exist for a particular data set an answerable one.

Section 5 presents a long-test viewpoint of test bias, making heavy use of Stout's theory of essential unidimensionality. No long-test test bias holding is defined. It is shown that if an equicontinuous balanced test score (a large class of reasonable to use test scores are such) displays appropriate convergence in probability behavior separately in each examinee group, then there can be no long-test test bias. Essential unidimensionality ($d_E = 1$) of a test with target ability as the latent trait is shown to exclude long-test test bias. Because one can statistically test for essential unidimensionality (Stout, 1987), this is a potentially very useful result. Theorem 5.3 is important as the long-test analogue to Theorem 4.2. It links potential for bias and scoring method to the existence of long-test test bias.

A long-test viewpoint of subtest validity is also present in Section 5. Informally stated, the main result is that $d_E = 1$ for a subtest with the latent trait being target ability implies subtest validity for all equicontinuous balanced scoring methods.

Section 6 considers test bias aggregated over target ability. The important concepts of unidirectional and bidirectional test bias are introduced. The relationship between differing discriminations across group and bidirectional test bias is explicated.

It is hoped that the above theory of test validity proves useful to theoreticians and practitioners alike.

Acknowledgement. The authors found discussions with Terry Ackerman, Paul Holland, Lloyd Humphreys, Kumar Joagdev, Brian Junker, Ratna Nandakumar, and Mark Reckase extremely useful in conducting the research above.

References

Ansley, T.N. and Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, *9*, 37–48.

Cleary, T. A. (1968). Test bias: prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement*, *5*, 115–124.

Cotter, D. E. and Berk R. A. (1981). Item bias in the WISC-R using black, white, and Hispanic learning disabled children, paper presented at the annual meeting of the *American Educational Research Association*, Los Angeles, CA, April.

Dorans, N. J. and Kulick E. (1983). Assessing unexpected differential item performance of Oriental candidates on SAT Form CSA6 and TSWE Form E33: November 1980 Administration, Unpublished Statistical Report No. SR-83-106, Educational Testing Service: Princeton, New Jersey.

Esary, J. D., Proschan, F. and Walkup, D.W. (1967). Association of random variables, with applications. *Annals of Mathematical Statistics*, *38*, 1466–1474.

Hambleton, R. K. and Swaminathan H. (1985). *Item Response Theory: Principles and Applications*, Kluwer-Nijhoff Publishing, Boston.

Holland, P. W. and Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. A chapter in *Test Validity*, Warner, H., and Braun, H.I. ed., Lawrence Erlbaum Associates, Hillsdale, NJ.

Hulin, C. L., Drasgow, F. and Parsons, C. K. (1983). *Item Response Theory: Applications to Psychological Measurement*, Dow Jones-Irwin, Homewood, IL.

Humphreys, L. (1984). A th ·tical and empirical study of the psychometric assessment of psychological test dimensionality and bias. ONR Research Proposal.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the predicting context. *Journal of Applied Psychology*, *71*, 327–333.

Ironson, G., Homan, S. , Willis, R. and Signer, B. (1984). The validity of item bias techniques with math word problems. *Applied Psychological Measurement*, *8*, 391–396.

Junker, B. (1989a). Essential Independence and likelihood based ability estimation of polytomous items. Submitted for publication.

Junker, B. (1989b). Conditional association, essential independence, and local independence. Submitted for publication.

Linn, R. L., Levine, M. V., Hastings, C. N. and Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement*, *5*, 159–173.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Mellenbergh, G. J. (1983). *Conditional item bias methods*, Plenum Publishing Corporation.

Reckase, M.D., Carlson, J. E., Ackerman, T. A. and Spray, J. A. (1986, June). *The interpretation of unidimensional IRT parameters when estimated from multidimensional data.* Paper presented at the annual meeting of the Psychometric Society, Toronto.

Rosenbaum, P. (1985). Comparing distributions of item responses for two groups. British Journal of Mathematical and Statistical Psychology, 38, 206–215.

Roznowski, M. (1987). Use of tests manifesting sex differences as measures of intelligence: implications for measurement bias. *Journal of Applied Psychology*, 72, 480–483.

Serfling, R. (1980). *Approximations Theorems of Mathematical Statistics*, John Wiley, New York.

Shealy, R. (1989). An item response theory based statistical procedure for detecting concurrent internal bias in ability tests. University of Illinois, Department of Statistics doctoral thesis.

Shealy, R. and Stout, W. (1990). A statistical test of psychological test bias. Submitted for publication.

Shepard, L. (1982). *Definitions of bias.* A chapter in *Handbook of Methods for Detecting Test Bias*, Berk, R.A., ed. Johns Hopkins University Press, Baltimore, Maryland.

Stanley, J. C. and Porter, A. C. (1967). Correlation of Scholastic Aptitude Test scores with college grades for Negroes versus whites. *Journal of Educational Measurement*, *4*, 199–218.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589–617.

Stout, W. (1989). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. To appear in *Psychometrika* in 1990.

Temp, G. (1971). Test bias: validity of the SAT for blacks and whites in thirteen integrated institutions. *Journal of Educational Measurement*, *8*, 245–251.

Thissen, D., Steinberg, L. and Wainer, H. (1988). *Use of item response theory in the study of group differences in trace lines*, a chapter in *Test Validity*, Wainer, H. and Braun, H. I. eds. Lawrence Erlbaum Associates, Hillsdale, NJ.

Wright, B. D., Mead, R. J. and Draba, R. E. (1976). *Detecting and correcting test item bias with a logistic response model.* Research Memorandum 22, University of Chicago, Department of Education.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125–145.

Dr. Terry Ackerman
Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. James Algina
1403 Norman Hall
University of Florida
Gainesville, FL 32605

Dr. Erling B. Andersen
Department of Statistics
Studiestraede 6
1455 Copenhagen
DENMARK

Dr. Ronald Armstrong
Rutgers University
Graduate School of Management
Newark, NJ 07102

Dr. Eva L. Baker
UCLA Center for the Study
of Evaluation
145 Moore Hall
University of California
Los Angeles, CA 90024

Dr. Laura L. Barnes
College of Education
University of Toledo
2801 W. Bancroft Street
Toledo, OH 43606

Dr. William M. Bart
University of Minnesota
Dept. of Educ. Psychology
330 Burton Hall
178 Pillsbury Dr., S.E.
Minneapolis, MN 55455

Dr. Isaac Bejar
Law School Admissions
Services
P.O. Box 40
Newtown, PA 18940-0040

Dr. Ira Bernstein
Department of Psychology
University of Texas
P.O. Box 19528
Arlington, TX 76019-0528

Dr. Menucha Birenbaum
School of Education
Tel Aviv University
Ramat Aviv 69978
ISRAEL

Dr. Arthur S. Blaiwes
Code N712
Naval Training Systems Center
Orlando, FL 32813-7100

Dr. Bruce Bloxom
Defense Manpower Data Center
99 Pacific St.
Suite 155A
Monterey, CA 93943-3231

Cdt. Arnold Bohrer
Sectie Psychologisch Onderzoek
Rekrutering-En Selectiecentrum
Kwartier Koningen Astrid
Bruijnstraat
1120 Brussels, BELGIUM

Dr. Robert Breaux
Code 281
Naval Training Systems Center
Orlando, FL 32826-3224

Dr. Robert Brennan
American College Testing
Programs
P. O. Box 168
Iowa City, IA 52243

Dr. Gregory Candell
CTB/McGraw-Hill
2500 Garden Road
Monterey, CA 93940

Dr. John B. Carroll
409 Elliott Rd., North
Chapel Hill, NC 27514

Dr. John M. Carroll
IBM Watson Research Center
User Interface Institute
P.O. Box 704
Yorktown Heights, NY 10598

Dr. Robert M. Carroll
Chief of Naval Operations
OP-01B2
Washington, DC 20350

Dr. Raymond E. Christal
UES LAMP Science Advisor
AFHRL/MOEL
Brooks AFB, TX 78235

Mr. Hua Hua Chung
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Dr. Norman Cliff
Department of Psychology
Univ. of So. California
Los Angeles, CA 90089-1061

Director, Manpower Program
Center for Naval Analyses
4401 Ford Avenue
P.O. Box 16268
Alexandria, VA 22302-0268

Director,
Manpower Support and
Readiness Program
Center for Naval Analysis
2000 North Beauregard Street
Alexandria, VA 22311

Dr. Stanley Collyer
Office of Naval Technology
Code 222
800 N. Quincy Street
Arlington, VA 22217-5000

Dr. Hans F. Crombag
Faculty of Law
University of Limburg
P.O. Box 616
Maastricht
The NETHERLANDS 6200 MD

Ms. Carolyn R. Crone
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

Dr. Timothy Davey
American College Testing Program
P.O. Box 168
Iowa City, IA 52243

Dr. C. M. Dayton
Department of Measurement
Statistics & Evaluation
College of Education
University of Maryland
College Park, MD 20742

Dr. Ralph J. DeAyala
Measurement, Statistics,
and Evaluation
Benjamin Bldg., Rm. 4112
University of Maryland
College Park, MD 20742

Dr. Lou DiBello
CERL
University of Illinois
103 South Mathews Avenue
Urbana, IL 61801

Dr. Dattprasad Divgi
Center for Naval Analysis
4401 Ford Avenue
P.O. Box 16268
Alexandria, VA 22302-0268

Mr. Hei-Ki Dong
Bell Communications Research
Room PYA-1K207
P.O. Box 1320
Piscataway, NJ 08855-1320

Dr. Fritz Drasgow
University of Illinois
Department of Psychology
603 E. Daniel St.
Champaign, IL 61820

Defense Technical
Information Center
Cameron Station, Bldg 5
Alexandria, VA 22314
(2 Copies)

Dr. Stephen Dunbar
224B Lindquist Center
for Measurement
University of Iowa
Iowa City, IA 52242

Dr. James A. Earles
Air Force Human Resources Lab
Brooks AFB, TX 78235

Dr. Susan Embretson
University of Kansas
Psychology Department
426 Fraser
Lawrence, KS 66045

Dr. George Englehard, Jr.
Division of Educational Studies
Emory University
210 Fishburne Bldg.
Atlanta, GA 30322

ERIC Facility-Acquisitions
2440 Research Blvd, Suite 550
Rockville, MD 20850-3238

Dr. Benjamin A. Fairbank
Operational Technologies Corp.
5825 Callaghan, Suite 225
San Antonio, TX 78228

Dr. Marshall J. Farr, Consultant
Cognitive & Instructional Sciences
2520 North Vernon Street
Arlington, VA 22207

Dr. P-A. Federico
Code 51
NPRDC
San Diego, CA 92152-6800

Dr. Leonard Feldt
Lindquist Center
for Measurement
University of Iowa
Iowa City, IA 52242

BEST COPY AVAILABLE

Dr. Richard L. Ferguson
American College Testing
P.O. Box 168
Iowa City, IA 52243

Dr. Gerhard Fischer
Liebiggasse 5/3
A 1010 Vienna
AUSTRIA

Dr. Myron Fischl
U.S. Army Headquarters
DAPE-MRR
The Pentagon
Washington, DC 20310-0300

Prof. Donald Fitzgerald
University of New England
Department of Psychology
Armidale, New South Wales 2351
AUSTRALIA

Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Alfred R. Fregly
AFOSR/NL, Bldg. 410
Bolling AFB, DC 20332-6448

Dr. Robert D. Gibbons
Illinois State Psychiatric Inst.
Rm 529W
1601 W. Taylor Street
Chicago, IL 60612

Dr. Janice Gifford
University of Massachusetts
School of Education
Amherst, MA 01003

Dr. Drew Gitomer
Educational Testing Service
Princeton, NJ 08541

Dr. Robert Glaser
Learning Research
& Development Center
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15260

Dr. Sherrie Gott
AFHRL/MOMJ
Brooks AFB, TX 78235-5601

Dr. Bert Green
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

Michael Habon
DORNIER GMBH
P.O. Box 1420
D-7990 Friedrichshafen 1
WEST GERMANY

Prof. Edward Haertel
School of Education
Stanford University
Stanford, CA 94305

Dr. Ronald K. Hambleton
University of Massachusetts
Laboratory of Psychometric
and Evaluative Research
Hills South, Room 152
Amherst, MA 01003

Dr. Delwyn Harnisch
University of Illinois
51 Gerty Drive
Champaign, IL 61820

Dr. Grant Henning
Senior Research Scientist
Division of Measurement
Research and Service
Educational Testing Service
Princeton, NJ 08541

Ms. Rebecca Hetter
Navy Personnel R&D Center
Code 63
San Diego, CA 92152-6800

Dr. Thomas M. Hirsch
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. Paul W. Holland
Educational Testing Service, 21-T
Rosedale Road
Princeton, NJ 08541

Dr. Paul Horst
677 G Street, #184
Chula Vista, CA 92010

Ms. Julia S. Hough
Cambridge University Press
40 West 20th Street
New York, NY 10011

Dr. William Howell
Chief Scientist
AFHRL/CA
Brooks AFB, TX 78235-5601

Dr. Lloyd Humphreys
University of Illinois
Department of Psychology
603 East Daniel Street
Champaign, IL 61820

Dr. Steven Hunka
3-104 Educ. N.
University of Alberta
Edmonton, Alberta
CANADA T6G 2G5

Dr. Huynh Huynh
College of Education
Univ. of South Carolina
Columbia, SC 29208

Dr. Robert Jannarone
Elec. and Computer Eng. Dept.
University of South Carolina
Columbia, SC 29208

Dr. Kumar Joag-dev
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright Street
Champaign, IL 61820

Dr. Douglas H. Jones
1280 Woodfern Court
Toms River, NJ 08753

Dr. Brian Junker
Carnegie-Mellon University
Department of Statistics
Schenley Park
Pittsburgh, PA 15213

Dr. Michael Kaplan
Office of Basic Research
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Dr. Milton S. Katz
European Science Coordination
Office
U.S. Army Research Institute
Box 65
FPO New York 09510-1500

Prof. John A. Keats
Department of Psychology
University of Newcastle
N.S.W. 2308
AUSTRALIA

Dr. Jwa-keun Kim
Department of Psychology
Middle Tennessee State
University
P.O. Box 522
Murfreesboro, TN 37132

Mr. Soon-Hoon Kim
Computer-based Education
Research Laboratory
University of Illinois
Urbana, IL 61801

Dr. G. Gage Kingsbury
Portland Public Schools
Research and Evaluation Department
501 North Dixon Street
P. O. Box 3107
Portland, OR 97209-3107

Dr. William Koch
Box 7246, Mess. and Eval. Ctr.
University of Texas-Austin
Austin, TX 78703

Dr. Richard J. Koubek
Department of Biomedical
& Human Factors
139 Engineering & Math Bldg.
Wright State University
Dayton, OH 45435

Dr. Leonard Kroeker
Navy Personnel R&D Center
Code 62
San Diego, CA 92152-6800

Dr. Jerry Lehnus
Defense Manpower Data Center
Suite 400
1600 Wilson Blvd
Rosslyn, VA 22209

Dr. Thomas Leonard
University of Wisconsin
Department of Statistics
1210 West Dayton Street
Madison, WI 53705

Dr. Michael Levine
Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. Charles Lewis
Educational Testing Service
Princeton, NJ 08541-0001

Mr. Rodney Lim
University of Illinois
Department of Psychology
603 E. Daniel St.
Champaign, IL 61820

Dr. Robert L. Linn
Campus Box 249
University of Colorado
Boulder, CO 80309-0249

Dr. Robert Lockman
Center for Naval Analysis
4401 Ford Avenue
P.O. Box 16268
Alexandria, VA 22302-0268

Dr. Frederic M. Lord
Educational Testing Service
Princeton, NJ 08541

Dr. Richard Luecht
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. George B. Macready
Department of Measurement
Statistics & Evaluation
College of Education
University of Maryland
College Park, MD 20742

Dr. Gary Marco
Stop 31-E
Educational Testing Service
Princeton, NJ 08451

Dr. Clessen J. Martin
Office of Chief of Naval
Operations (OP 13 F)
Navy Annex, Room 2832
Washington, DC 20350

Dr. James R. McBride
HumRRO
6430 Elmhurst Drive
San Diego, CA 92120

Dr. Clarence C. McCormick
HQ, USMEPCOM/MEPCT
2500 Green Bay Road
North Chicago, IL 60064

Mr. Christopher McCusker
University of Illinois
Department of Psychology
603 E. Daniel St.
Champaign, IL 61820

Dr. Robert McKinley
Educational Testing Service
Princeton, NJ 08541

Mr. Alan Mead
c/o Dr. Michael Levine
Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. Timothy Miller
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. Robert Mislevy
Educational Testing Service
Princeton, NJ 08541

Dr. William Montague
NPRDC Code 13
San Diego, CA 92152-6800

Ms. Kathleen Moreno
Navy Personnel R&D Center
Code 62
San Diego, CA 92152-6800

Headquarters Marine Corps
Code MPI-20
Washington, DC 20380

Dr. Ratna Nandakumar
Educational Studies
Willard Hall, Room 213E
University of Delaware
Newark, DE 19716

Library, NPRDC
Code P201L
San Diego, CA 92152-6800

Librarian
Naval Center for Applied Research
in Artificial Intelligence
Naval Research Laboratory
Code 5510
Washington, DC 20375-5000

Dr. Harold F. O'Neil, Jr.
School of Education - WPH 801
Department of Educational
Psychology & Technology
University of Southern California
Los Angeles, CA 90089-0031

Dr. James B. Olsen
WICAT Systems
1875 South State Street
Orem, UT 84058

Office of Naval Research,
Code 1142CS
800 N. Quincy Street
Arlington, VA 22217-5000
(6 Copies)

Dr. Judith Orasanu
Basic Research Office
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Dr. Jesse Orlansky
Institute for Defense Analyses
1801 N. Beauregard St.
Alexandria, VA 22311

Dr. Peter J. Pashley
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

Wayne M. Patience
American Council on Education
GED Testing Service, Suite 20
One Dupont Circle, NW
Washington, DC 20036

Dr. James Paulson
Department of Psychology
Portland State University
P.O. Box 751
Portland, OR 97207

Dept. of Administrative Sciences
Code 54
Naval Postgraduate School
Monterey, CA 93943-5026

Dr. Mark D. Reckase
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. Malcolm Ree
AFHRL/MOA
Brooks AFB, TX 78235

Mr. Steve Reise
N660 Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455-0344

Dr. Carl Ross
CNET-PDCD
Building 90
Great Lakes NTC, IL 60088

Dr. J. Ryan
Department of Education
University of South Carolina
Columbia, SC 29208

Dr. Fumiko Samejima
Department of Psychology
University of Tennessee
310B Austin Peay Bldg.
Knoxville, TN 37916-0900

Mr. Drew Sands
NPRDC Code 62
San Diego, CA 92152-6800

Lowell Schoer
Psychological & Quantitative
Foundations
College of Education
University of Iowa
Iowa City, IA 52242

Dr. Mary Schratz
4100 Parkside
Carlsbad, CA 92008

Dr. Dan Segall
Navy Personnel R&D Center
San Diego, CA 92152

Dr. Robin Shealy
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Dr. Kazuo Shigemasu
7-9-24 Kugenuma-Kaigan
Fujisawa 251
JAPAN

Dr. Randall Shumaker
Naval Research Laboratory
Code 5510
4555 Overlook Avenue, S.W.
Washington, DC 20375-5000

Dr. Richard E. Snow
School of Education
Stanford University
Stanford, CA 94305

Dr. Richard C. Sorensen
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Judy Spray
ACT
P.O. Box 168
Iowa City, IA 52243

Dr. Martha Stocking
Educational Testing Service
Princeton, NJ 08541

Dr. Peter Stoloff
Center for Naval Analysis
4401 Ford Avenue
P.O. Box 16268
Alexandria, VA 22302-0268

Dr. William Stout
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Dr. Hariharan Swaminathan
Laboratory of Psychometric and
Evaluation Research
School of Education
University of Massachusetts
Amherst, MA 01003

Mr. Brad Sympson
Navy Personnel R&D Center
Code-62
San Diego, CA 92152-6800

Dr. John Tangney
AFOSR/NL, Bldg. 410
Bolling AFB, DC 20332-6448

Dr. Kikumi Tatsuoka
Educational Testing Service
Mail Stop 03-T
Princeton, NJ 08541

Dr. Maurice Tatsuoka
Educational Testing Service
Mail Stop 03-T
Princeton, NJ 08541

Dr. David Thissen
Department of Psychology
University of Kansas
Lawrence, KS 66044

Mr. Thomas J. Thomas
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

Mr. Gary Thomasson
University of Illinois
Educational Psychology
Champaign, IL 61820

Dr. Robert Tsutakawa
University of Missouri
Department of Statistics
222 Math. Sciences Bldg.
Columbia, MO 65211

Dr. Ledyard Tucker
University of Illinois
Department of Psychology
603 E. Daniel Street
Champaign, IL 61820

Dr. David Vale
Assessment Systems Corp.
2233 University Avenue
Suite 440
St. Paul, MN 55114

Dr. Frank L. Vicino
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Howard Wainer
Educational Testing Service
Princeton, NJ 08541

Dr. Michael T. Waller
University of Wisconsin-Milwaukee
Educational Psychology Department
Box 413
Milwaukee, WI 53201

Dr. Ming-Mei Wang
Educational Testing Service
Mail Stop 03-T
Princeton, NJ 08541

Dr. Thomas A. Warm
FAA Academy AAC934D
P.O. Box 25082
Oklahoma City, OK 73125

Dr. Brian Waters
HumRRO
1100 S. Washington
Alexandria, VA 22314

Dr. David J. Weiss
N660 Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455-0344

Dr. Ronald A. Weitzman
Box 146
Carmel, CA 93921

Major John Welsh
AFHRL/MOAN
Brooks AFB, TX 78223

Dr. Douglas Wetzel
Code 51
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Rand R. Wilcox
University of Southern
California
Department of Psychology
Los Angeles, CA 90089-1061

German Military Representative
ATTN: Wolfgang Wildgrube
Streitkraefteamt
D-5300 Bonn 2
4000 Brandywine Street, NW
Washington, DC 20016

Dr. Bruce Williams
Department of Educational
Psychology
University of Illinois
Urbana, IL 61801

Dr. Hilda Wing
Federal Aviation Administration
800 Independence Ave, SW
Washington, DC 20591

Mr. John H. Wolfe
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. George Wong
Biostatistics Laboratory
Memorial Sloan-Kettering
Cancer Center
1275 York Avenue
New York, NY 10021

Dr. Wallace Wulfeck, III
Navy Personnel R&D Center
Code 51
San Diego, CA 92152-6800

Dr. Kentaro Yamamoto
02-T
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

Dr. Wendy Yen
CTB/McGraw Hill
Del Monte Research Park
Monterey, CA 93940

Dr. Joseph L. Young
National Science Foundation
Room 320
1800 G Street, N.W.
Washington, DC 20550

Mr. Anthony R. Zara
National Council of State
Boards of Nursing, Inc.
625 North Michigan Avenue
Suite 1544
Chicago, IL 60611